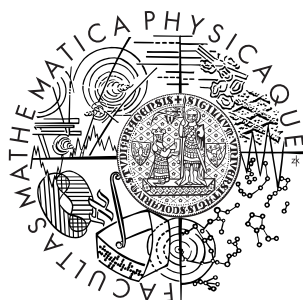


Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCA



Vlasta Poliačková

Vliv chyb v modelu regrese

KPMS

Vedoucí diplomové práce: doc. RNDr. Petr Lachout, CSc.
Studijní program: Pravděpodobnost, matematická statistika a
ekonometrie

2012

Na tomto mieste by som rada poďakovala predovšetkým vedúcemu mojej diplomovej práce doc. RNDr. Petrovi Lachoutovi, CSc. za cenné rady a čas, ktorý mi venoval, ďalej by som sa chcela poďakovať Janovi Majdanovi za osobnú podporu a pomoc pri revízii slovenčiny v texte.

Prehlasujem, že som túto diplomovú prácu vypracovala samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, obzvlášť skutočnosť, že Univerzita Karlova v Praze má právo na uzatvorenie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

V Prahe dňa 6.12.2012

Vlasta Poliačková

Názov práce: Vliv chyb v modelu regrese

Autor: Bc. Vlasta Poliačková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúcí diplomovej práce: doc. RNDr. Petr Lachout, CSc.

e-mail vedúceho: Petr.Lachout@mff.cuni.cz

Abstrakt: Predložená práca sa zaoberá regresným modelom a vplyvom chýb v regresii. Popisuje rôzne typy porušení predpokladov kladených na chybový člen a ich vplyv na vlastnosti regresného modelu. Následne sú v texte rozobraté štatistické prístupy aplikovateľné v prípade porušenia predpokladov regresného modelu ako sú heteroskedasticita alebo autokorelovanosť reziduálnej zložky. V aplikačnej časti sú využité hlavne poznatky z Boxovej–Jenkinsovej metodológie. V tejto časti je podrobne popísaný postup budovania Boxových–Jenkinsových modelov na rôznych reálnych finančných časových radoch, ku ktorým sú následne vytvorené predikcie budúcich hodnôt. Pri spracovaní dát sú použité modely typu ARMA, AIRMA a SARIMA. Na príklade sú predpovede porovnané s reálnymi budúcimi hodnotami radu.

Kľúčové slová: regresia, porušenie predpokladov, reziduálna zložka, Boxova-Jenkinsova metodológia, časový rad

Title: Influence of errors to regression model

Author: Bc. Vlasta Poliačková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Petr Lachout, CSc.

Supervisor's e-mail address: Petr.Lachout@mff.cuni.cz

Abstract: The submitted work deals with the regression model, and the influence of errors to regression. Thesis describes different types of violations of assumptions required to the error term and their impact to the properties of the regression model. In the next part, there are discussed various statistical approaches applicable in the case of violation assumptions of regression model such as heteroscedasticity or autocorrelation of the residuals. In the application part, there is used mainly knowledge of Box - Jenkins methodology. In this section it is described in detail how to build a Box - Jenkins models and forecasts of future values for various real financial time series. In processing of the data are used models of ARMA, ARIMA and SARIMA. In an example, forecasts of the models are compared to real future values of the time series.

Keywords: regression, violation of assumptions, error term, Box-Jenkins methodology, time series

Obsah

1	Úvod	6
2	Klasický a zobecnený model lineárnej regresie	8
2.1	Metóda najmenších štvorcov a Gaussova–Markovova veta	9
2.2	Zovšeobecnená metóda najmenších štvorcov a Aitkenova veta	12
2.3	Bodové odhady a testovanie hypotéz na malých výberoch	14
2.3.1	Maximálne vierohodný odhad	14
2.3.2	Odhady na malých výberoch	15
2.3.3	Testovanie hypotéz	17
2.3.4	Odhady a testovanie hypotéz na veľkých výberoch	20
3	Porušenie základných predpokladov	24
3.1	„Feasible“ zobecnená metóda najmenších štvorcov	24
3.1.1	Definícia „Feasible“ GLM	25
3.1.2	Vlastnosti odhadu „feasible“ zovšeobecnenou metódou najmenších štvorcov	26
3.1.3	„Zdanlivo“ nesúvisiace regresie	29
3.1.4	Odhady: asymptotické výsledky	33
3.1.5	Zdanlivo nesúvisiace regresie: výsledky na malom výbere	34
3.2	Heteroskedasticita	36
3.2.1	Odhady v rôznych formách heteroskedasticity	37
3.2.2	Detekcia heteroskedasticity	45
3.2.3	Test založený na pomere vierohodností pre skupinovú heteroskedasticitu	45
3.2.4	Test heteroskedasticity ako funkcie exogénnych premenných	46
3.2.5	Chowov test	52
3.3	Autokorelácia	54
3.3.1	Autokorelačné vlastnosti časových radov	54
3.3.2	Autoregresná postupnosť prvého rádu AR(1)	57
3.3.3	Odhad na malom výbere v AR(1)	62

3.3.4	ARMA procesy v reziduálnej zložke regresného modelu	63
3.3.5	Procesy vykazujúce nekonštantný trend, zmenu úrovne alebo zmenu trendu	67
3.3.6	Procesy so sezónnou zložkou	68
3.3.7	Predpovede v Boxovej–Jenkinsovej metodológii	69
3.3.8	Autoregresné modely volatility	69
4	Aplikácia na reálnych dátach	71
4.1	Popis dát	71
4.2	Spracovanie a analýza dát	72
4.2.1	Príklad 1	72
4.2.2	Príklad 2	92
4.2.3	Príklad 3	96
5	Záver	102
	Literatúra	103
	Zoznam tabuliek	104
	Zoznam obrázkov	106

Kapitola 1

Úvod

Regresná analýza patrí k najvýznamnejším metódam matematickej štatistiky a má veľmi široké použitie. Preto je veľmi často používaným štatistickým nástrojom. Ako prvý použil pojem regresie vo význame, akým je dnes tento pojem chápaný v matematickej štatistike, Francis Galton, keď v 19. storočí skúmal závislosť výšky potomkov na výške rodičov. Zistil pritom, že synovia vysokých otcov sú zvyčajne nižší ako ich otcovia a naopak synovia nízkych otcov sú obvykle vyšší než ich otcovia. Teda išlo o spätný posun, čiže regresiu, k priemeru (viď [12]).

Úlohou regresie je vysvetliť zmeny hodnôt jednej premennej, ktorá sa zvyčajne nazýva *vysvetľovaná premenná*, resp. *závislá premenná*, resp. *odozva*, zmenami hodnôt iných premenných nazývaných *vysvetľujúce premenné*, resp. *nezávislé premenné*, resp. *regresory*. Pri praktickom použití regresného modelu a pri odhadovaní jeho parametrov má zásadný vplyv *chybový člen* (*náhodná*, resp. *reziduálna zložka*). Klasický model predpokladá, že reziduá sú nezávislé a rovnako rozdelené (*i.i.d.* - *independent identically distributed*). Tento predpoklad je ale na reálnych dátach pomerne často porušený. Je preto nutné zvážiť vhodné zoslabenie predpokladu. Napríklad môžeme vyžadovať, aby chyby tvorili biely šum, slabo alebo silno stacionárny proces, prípadne nestacionárny proces, či proces s meniacim sa rozptylom, a pod. V prípade nesplnenia základných predpokladov klasického regresného modelu, môžu byť samozrejme ovplyvnené aj vlastnosti modelu. V tomto texte uvedieme základnú terminológiu týkajúcu sa regresie, zhrnieme rôzne možnosti zovšeobecnenia modelu pri porušení predpokladov klasického modelu a dopad na vlastnosti modelu. V ďalšom sa zameriame na procesy, ktoré sa používajú ako chybová zložka v regresných modeloch. Tieto procesy sú často používané pri modelovaní finančných časových radov, čo ukážeme aj na reálnych príkladoch. Regresii je síce venované množstvo štatistickej literatúry, avšak cieľom tejto práce je popísať možnosti riešenia v prípade porušenia predpokladov o chybovej zložke a hlavne ukážka aplikácie niektorých týchto postupov na reálnych dátach.

Nakoľko sa budeme zaoberať najmä porušeniami predpokladov regresného modelu, v nasledujúcej kapitole je uvedená definícia klasického lineárneho regresného modelu a zhrnutie základných teoretických poznatkov o regresných modeloch. Kapitola 3 je zameraná na rôzne typy porušenia predpokladov klasického regresného modelu a možnosti výstavby modelu v prítomnosti týchto porušení. Ďalej stojí

za zmienku, že v tejto kapitole sú tiež zadefinované jednotlivé procesy, ktoré sa používajú k modelovaniu reziduálnej zložky v regresnom modele. Teoretické poznatky sú prebraté z uvedeného zoznamu literatúry. V kapitole 4 je popísaná konkrétna aplikácia autoregresných procesov na reálnych finančných dátach, presnejšie, budeme analyzovať a modelovať tri rôzne reálne časové rady, pričom všetky tri časové rady popisujú stav zostatkov na bankových účtoch (či už stav depozitných zostatkov alebo stav pohľadávok) na nejakom výbere klientov. A konečne, v kapitole 5 sú zhrnuté závery, ku ktorým som počas písania tejto práce dospela.

Kapitola 2

Klasický a zobecnený model lineárnej regresie

Najjednoduchším typom medzi regresnými modelmi je lineárna regresia. Slovo *lineárna* sa v tomto prípade vzťahuje k parametrom modelu, t.j regresný model má byť až na chybový člen lineárnou funkciou parametrov modelu - regresných koeficientov. Môžeme tiež povedať, že stredná hodnota vysvetľovanej premennej je lineárnou funkciou regresných koeficientov. Nakoľko v nasledujúcom texte budeme pracovať s vektorovým zápisom regresného modelu, uvedieme tiež definíciu vo vektorovom zápise rovnako ako je uvedená v [7].

Definícia 2.1 *Klasický model lineárnej regresie je definovaný ako*

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}, \quad (2.1)$$

kde $\mathbf{y}' = (y_1, y_2, \dots, y_T)$, $\mathbf{e}' = (e_1, e_2, \dots, e_T)$, $\beta' = (\beta_1, \beta_2, \dots, \beta_T)$ a

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ \vdots & \vdots & \dots & \vdots \\ x_{T1} & x_{T2} & \dots & x_{TK} \end{pmatrix}$$

a platia nasledujúce predpoklady:

- \mathbf{X} je nenáhodná matica s hodnotou $K \leq T$, kde T je rozsah výberu a pre $T \rightarrow \infty$ platí $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X}/T) = \mathbf{Q}$, kde \mathbf{Q} je konečná regulárna matica
- Pre vektor náhodných chýb \mathbf{e} platí: $\mathbf{E}\mathbf{e} = \mathbf{0}$ a $\mathbf{E}\mathbf{e}\mathbf{e}' = \sigma^2\mathbf{I}$, kde $\mathbf{E}(\cdot)$ je stredná hodnota, t.j. náhodné chyby sú i.i.d. a σ^2 je spravidla neznámy parameter.

Keďže v klasickom modeli lineárnej regresie platí, že $\mathbf{E}\mathbf{e} = \mathbf{0}$, vidíme, že $\mathbf{E}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta$, takže očakávané hodnoty náhodných premenných y_t sú lineárnou funkciou vysvetľujúcich premenných x_{t1}, \dots, x_{tK} .

Často hovoríme, že premenná y_t je vysvetľovaná premennými x_{t1}, \dots, x_{tK} .

2.1 Metóda najmenších štvorcov a Gaussova–Markovova veta

Ako už bolo spomenuté, pri regresnom modelovaní chceme popísať závislé premenné pomocou nezávislých premenných, pričom závislé aj nezávislé premenné sú vlastne pozorovania, ktoré máme k dispozícii. Takže v modele klasickej regresie potrebujeme odvodiť hodnoty pre neznáme parametre $\beta_1, \beta_2, \dots, \beta_K$ a σ^2 pre daný náhodný výber pozorovaní. Pre výpočet neznámych parametrov sa v klasickej regresii používa tzv. *metóda najmenších štvorcov* (*ordinary least squares - OLS*), ktorá minimalizuje súčet štvorcových chýb:

$$S = \sum_{t=1}^T (y - \beta_1 x_{t1} - \beta_2 x_{t2} - \dots - \beta_K x_{tK})^2 = (y - X\beta)'(y - X\beta).$$

V praxi to znamená, že hľadáme konkrétnu priamku, pre ktorú bude platiť, že súčet druhých mocnín vzdialeností skutočných pozorovaní od tejto priamky bude menší alebo rovný ako takýto súčet pre ľubovoľnú inú priamku. Samotný výpočet prebehne tak, že spočítame lokálne extrém, ktoré získame ako stacionárne body S ako funkcie v premennej β . T.j. S zderivujeme podľa β , a to, že sa jedná o lokálne minimum overíme tým, že druhá derivácia S podľa β je pozitívne definitná matica. V prípade lineárnej regresie teda dostaneme jeden stacionárny bod, pretože:

$$\frac{\partial S}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{0},$$

čo je ekvivalentné s tým, že odhad parametru β metódou najmenších štvorcov má tvar

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

pretože matica druhých derivácií \mathbf{S} , je $\partial^2 \mathbf{S} / \partial \beta \partial \beta' = 2\mathbf{X}\mathbf{X}'$, takže je pozitívne definitná (viď dôkaz pomocného tvrdenia v dôkaze Gaussovej-Markovovej vety), a teda jedná sa o minimum \mathbf{S} .

Ďalej budeme potrebovať ešte niekoľko bežne používaných definícií a tvrdení:

Definícia 2.2 *Nech β^* je odhad β , potom β^* je nestranný (tiež nevychýlený) práve vtedy, keď $E\beta^* = \beta$.*

Definícia 2.3 *Nestranný odhad je eficientný vzhľadom k triede odhadov práve vtedy, keď jeho rozptyl nie je väčší ako rozptyl ľubovoľného odhadu z tejto triedy.*

Takže z definície eficientného odhadu je zrejmé, že sa jedná o odhad, od ktorého nenájdeme v danej triede odhadov lepší odhad. Trieda odhadov však môže obsahovať kľudne viac ako jeden eficientný odhad.

Veta 2.1 Gaussova–Markovova veta. *Nech platia predpoklady klasického lineárneho modelu. Potom OLS-odhad $\hat{\beta}$ je nestranný a je to najlepší nestranný lineárny odhad (NNLO) v triede nestranných lineárnych odhadov.*

Dôkaz Substitúciou za \mathbf{y} v *OLS*-odhade $\boldsymbol{\beta}$ dostaneme

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}.$$

Takže $\hat{\boldsymbol{\beta}}$ je nestranný, pretože

$$\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}\mathbf{e} = \boldsymbol{\beta},$$

čo vyplýva z toho, že \mathbf{X} nie je stochastická a z predpokladu $\mathbb{E}\mathbf{e} = \mathbf{0}$. Potom kovariančná matica

$$\mathbb{E}(\hat{\boldsymbol{\beta}} - \mathbb{E}\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \mathbb{E}\hat{\boldsymbol{\beta}})' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathbf{e}\mathbf{e}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Aby sme ukázali, že *OLS*-odhad $\boldsymbol{\beta}$ je eficientný, uvažujme nejaký iný lineárny odhad $\boldsymbol{\beta}$, ktorý označíme napríklad $\boldsymbol{\beta}^*$ a pre ktorý platí

$$\boldsymbol{\beta}^* = \mathbf{H}\mathbf{y},$$

kde $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C}$, a \mathbf{C} je matica konštánt. Potom

$$\boldsymbol{\beta}^* = \mathbf{H}\mathbf{y} = \mathbf{H}\mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{e}$$

a $\mathbb{E}\boldsymbol{\beta}^* = \mathbf{H}\mathbf{X}\boldsymbol{\beta}$ a teda $\mathbf{H}\mathbf{X}$ musí byť jednotková matica kvôli nestrannosti. Dosadením ale zistíme, že

$$\mathbf{H}\mathbf{X} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} + \mathbf{C}\mathbf{X},$$

takže $\mathbf{H}\mathbf{X} = \mathbf{I}$ iba vtedy, keď $\mathbf{C}\mathbf{X} = \mathbf{0}$. Kovariančná matica $\boldsymbol{\beta}^*$ je teda

$$\begin{aligned} \text{var}\boldsymbol{\beta}^* &= \mathbb{E}(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})' \\ &= \mathbb{E}(\mathbf{H}\mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{e} - \boldsymbol{\beta})(\mathbf{H}\mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{e} - \boldsymbol{\beta})', \end{aligned}$$

pretože $\mathbf{H}\mathbf{X} = \mathbf{I}$ dostaneme

$$\begin{aligned} \text{var}\boldsymbol{\beta}^* &= \mathbb{E}(\mathbf{H}\mathbf{e}\mathbf{e}'\mathbf{H}') \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C}][\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{C}] \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C} + \mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{C}\mathbf{C}'] \end{aligned}$$

a pretože $\mathbf{C}\mathbf{X} = \mathbf{0}$,

$$\text{var}\boldsymbol{\beta}^* = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2\mathbf{C}\mathbf{C}'.$$

Ale $\mathbf{C}\mathbf{C}' \geq \mathbf{0}$, pretože $\mathbf{C}\mathbf{C}'$ je symetrická štvorcová matica a teda je pozitívne semidefinitná, takže $\boldsymbol{\beta}^*$ nemá menší rozptyl ako $\hat{\boldsymbol{\beta}}$. Pozitívna semidefinitnosť $\mathbf{C}\mathbf{C}'$ vyplýva z definície pozitívne semidefinitnej matice a z nasledujúceho pomocného tvrdenia.

Definícia 2.4 *Nech $\mathbf{A} = [a_{ij}]$ je symetrická matica konštánt typu $n \times n$ a \mathbf{x}' je vektor premenných typu $1 \times n$. Potom kvadratická forma $\mathbf{x}'\mathbf{A}\mathbf{x}$ je pozitívne semidefinitná práve vtedy, keď $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ pre všetky vektory \mathbf{x} . Ekvivalentne potom hovoríme, že matica \mathbf{A} je pozitívne semidefinitná.*

Tvrdenie *Nech \mathbf{C} je ľubovoľná matica typu $K \times T$. Potom matica $\mathbf{C}\mathbf{C}'$ je pozitívne semidefinítaná.*

Dôkaz tvrdenia. Súčin $\mathbf{C}\mathbf{C}'$ je zjavne symetrická štvorcová matica $K \times K$:

$$\mathbf{C}\mathbf{C}' = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1T} \\ c_{21} & c_{22} & \dots & c_{2T} \\ \vdots & \vdots & \dots & \vdots \\ c_{K1} & c_{K2} & \dots & c_{KT} \end{bmatrix} \begin{bmatrix} c_{11} & c_{21} & \dots & c_{K1} \\ c_{12} & c_{22} & \dots & c_{K2} \\ \vdots & \vdots & \dots & \vdots \\ c_{1T} & c_{2T} & \dots & c_{KT} \end{bmatrix} = \left(\sum_{t=1}^T c_{it}c_{jt} \right)_{i,j=1}^K.$$

Takže stačí ukázať, že pre každý vektor $\mathbf{x}' = (x_1, x_2, \dots, x_K)$ kvadratická forma $\mathbf{x}'(\mathbf{C}\mathbf{C}')\mathbf{x}$ je nezáporná. Pretože násobenie matíc je asociatívne, platí:

$$\begin{aligned} \mathbf{x}'(\mathbf{C}\mathbf{C}')\mathbf{x} &= (\mathbf{x}'\mathbf{C})(\mathbf{C}'\mathbf{x}) \\ &= \left(\sum_{t=1}^K x_i c_{i1}, \dots, \sum_{t=1}^K x_i c_{iK} \right) \left(\sum_{t=1}^K x_i c_{i1}, \dots, \sum_{t=1}^K x_i c_{iK} \right)' \\ &= \sum_{l=1}^T \left(\sum_{t=1}^K x_t c_{tl} \right)^2 \geq 0 \end{aligned}$$

Q.E.D.

Tým je dokázané, že OLS-odhad $\hat{\beta}$ je najlepší nestranný lineárny odhad v triede nestranných lineárnych odhadov.

Q.E.D.

Zostáva nám ešte získať odhad parametru σ^2 . V klasickom regresnom modeli. Pri splnení všetkých predpokladov, spočítame odhad parametru σ^2 veľmi jednoducho pomocou OLS-odhadu $\hat{\beta}$ tak, ako uvádza nasledujúca veta, ktorej dôkaz je napríklad v [7].

Veta 2.2 *Odhad parametru σ^2 v klasickom lineárnom regresnom modeli je $\hat{\sigma}^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}/(T - K)$, kde $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}$ a jedná sa o nestranný odhad parametru σ^2 .*

Teraz sa budeme chvíľu zaoberať veľkými výbermi a asymptotickými vlastnosťami distribúcie náhodného výberu alebo nejakého odhadu, ak rozsah výberu bude rásť do nekonečna.

Definícia 2.5 *Nech $\{z_T : T = 1, 2, \dots\}$ je postupnosť náhodných veličín. Hovoríme, že z_T konverguje v pravdepodobnosti ku konštante l (a píšeme $\text{plim } z_T = l$) práve vtedy, keď platí: $\forall \epsilon > 0 : \lim_{T \rightarrow \infty} P[|z_T - l| < \epsilon] = 1$.*

Definícia 2.6 *Nech $\hat{\theta}_T$ je odhad parametru θ . Ak $\hat{\theta}_T \rightarrow \theta$ v pravdepodobnosti, potom $\hat{\theta}_T$ je konzistentný odhad θ .*

Definícia 2.7 Hovoríme, že postupnosť náhodných veličín $\{z_T : T = 1, 2, \dots\}$, ktoré majú konečné druhé momenty, konverguje podľa kvadratického stredy k náhodnej veličine z , taktiež s konečnými druhými momentami, práve vtedy, keď

$$\lim_{n \rightarrow \infty} E|z_n - z|^2 = 0.$$

Veta 2.3 Konvergenca podľa kvadratického stredy: Odhad $\hat{\theta}_T$ je konzistentný odhad parametru θ ak je nevychýlený a jeho rozptyl asymptoticky „mizne“ (t.j. keď konverguje podľa kvadratického stredy).

Veta 2.4 Khintchine: Nech $z_t, t = 1, 2, \dots$ je postupnosť i.i.d. náhodných veličín s konečnou strednou hodnotou μ , potom $\text{plim } \bar{z}_t = \mu$.

Dôkazy k vetám 2.4 a 2.3 sú k dohľadaniu v [7].

2.2 Zovšeobecnená metóda najmenších štvorcov a Aitkenova veta

Predpoklad, že náhodné chyby sú i.i.d. je v praxi často porušený, a teda neudržateľný. Pomerne často porušovanými predpokladmi modelu klasickej regresie sú heteroskedasticita a korelovanosť chýb (t.j. porušenie predpokladu $E\mathbf{e}\mathbf{e}' = \sigma^2\mathbf{I}$). Čo znamená, že v praxi často nastáva, že chyby nemajú konštantné rozptyly alebo nie sú nekorelované. Uvažujme jednoduchý príklad. Pre pozorovania $t = 1, 2, \dots, T'$, je rozptyl chýb e_t rovný σ^2 , zatiaľ čo pre pozorovania $T' < t \leq T$ je rozptyl e_t rovný $c\sigma^2$, kde c je známa konštanta. V tomto prípade:

$$E\mathbf{e}\mathbf{e}' = \sigma^2 \begin{pmatrix} \mathbf{I}_{T'} & \mathbf{0} \\ \mathbf{0} & c\mathbf{I}_{T-T'} \end{pmatrix} = \sigma^2\mathbf{\Omega},$$

kde $\mathbf{I}_{T'}$ a $\mathbf{I}_{T-T'}$ sú jednotkové matice rádu T' a $T - T'$. Podobne môžu existovať rôzne kovariančné štruktúry rôzne od $\sigma^2\mathbf{I}$. Formálne:

Definícia 2.8 Zovšeobecnený model najmenších štvorcov (angl. skr. GLM - Generalized Least Squares Model) je definovaný ako

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kde \mathbf{y} je $T \times 1$ vektor pozorovaní závislej premennej, $\boldsymbol{\beta}$ je $K \times 1$ vektor neznámych koeficientov a

- \mathbf{X} je nenáhodná matica $T \times K$ pozorovaní vysvetľujúcich premenných s hodnotou K ,
- \mathbf{e} je $T \times 1$ vektor náhodných nepozorovateľných chýb, pre ktoré platí: $E\mathbf{e} = \mathbf{0}$ a $E\mathbf{e}\mathbf{e}' = \sigma^2\mathbf{\Omega}$, kde $\mathbf{\Omega}$ je známa symetrická pozitívne definitná matica a

- $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}/T) = \mathbf{Q}_{\mathbf{\Omega}}$ je konečná a regulárna,

teda pripúšťame korelované chyby a rôzne rozptyly

Veta 2.5 Aitken: Odhad parametru β zobecnenou metódou najmenších štvorcov, ktorý označíme $\tilde{\beta}$ je:

$$\tilde{\beta} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}$$

a je to eficientný lineárny a nevychýlený odhad parametru β .

Dôkaz Aitkenovej vety je tiež uvedený v [7].

Vlastnosti odhadov metódou najmenších štvorcov pri nedodržaní podmienok klasickej regresie:

Komplexné skúmanie splnenia predpokladov kladených na chybovú zložku je obvykle zložitá úloha. Naviac použitie bežnej metódy najmenších štvorcov (*Ordinary Least Squares*, resp. *OLS*) je zvyčajne jednoduchšie ako použitie zobecnenej metódy najmenších štvorcov (*Generalized Least Squares*, resp. *GLS*). Preto v mnohých prípadoch dochádza k nevhodnému použitiu *OLS* miesto *GLS*.

Uvažujme model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e},$$

kde $\mathbf{E}\mathbf{e}\mathbf{e}' = \sigma^2\mathbf{\Omega}$ a $\mathbf{\Omega}$ je pozitívne definitná a symetrická matica, ale $\mathbf{\Omega} \neq \mathbf{I}$ potom:

- *OLS* odhad $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ je nestranný. Ak $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X}/T)$ je konečná a regulárna a $\lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{\Omega}\mathbf{X}/T)$ je konečná, $\hat{\beta}$ je konzistentný.

Plynie to z nasledujúcich faktov:

- $\mathbf{E}\hat{\beta} = \beta + \mathbf{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} = \beta$
- $\text{plim } \hat{\beta} = \beta + \lim_{T \rightarrow \infty} (\mathbf{X}'\mathbf{X}/T)^{-1} \text{plim } \mathbf{X}'\mathbf{e}/T$
- $\mathbf{E}\mathbf{X}'\mathbf{e}/T = 0$
- $\text{var}\mathbf{X}'\mathbf{e}/T = \sigma^2(\mathbf{X}'\mathbf{\Omega}\mathbf{X}/T^2)$

- Kovariančná matica $\hat{\beta}$ je $\mathbf{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.
- Pri chybnej aplikácii *OLS* miesto *GLS* je rozptyl minimálne taký veľký, alebo väčší ako pri *GLS*

Nesprávne použitie metódy najmenších štvorcov vedie k neplatnému testovaniu hypotéz, pretože skutočná kovariančná matica $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ sa nahradí $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$. Z predchádzajúcich vlastností pre $\mathbf{\Omega} \neq \mathbf{I}$ pri použití $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ vzniká vychýlenie, pretože $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \neq (\mathbf{X}'\mathbf{X})^{-1}$ a $\mathbf{E}\hat{\sigma}^2 \neq \sigma^2$. Táto zámiena môže ľahko spôsobiť, že dôjde k nesprávnym štatistickým záverom.

2.3 Bodové odhady a testovanie hypotéz na malých výberoch

Pri malých výberoch sú kladené silnejšie predpoklady na distribúciu chybového člena ako sme boli zvyknutí doteraz. V tejto sekcii budeme okrem predpokladov z predchádzajúcich častí navyše ďalej predpokladať, že chyby majú normálne rozdelenie. Takýto model nazývame *klasický normálny lineárny regresný model*. Tento model pripúšťa odhadovanie parametrov pomocou metódy maximálnej vierohodnosti a konštrukciu testovania hypotéz pomerom vierohodnosti.

2.3.1 Maximálne vierohodný odhad

Ďalšou metódou, ktorá sa pomerne často používa k odhadnutiu regresných koeficientov je metóda maximálnej vierohodnosti. Maximálnu vierohodnosť môžeme použiť v prípadoch, keď poznáme stochastickú distribúciu pozorovaní, v našom prípade distribúciu reziduí.

Definícia 2.9 *Nech $x_t, t = 1, 2, \dots, T$ je náhodný výber z populácie charakterizovanej hustotou $f(x_t; \theta)$, ktorá závisí na parametre $\theta \in \Theta$, kde Θ je parametrický priestor. Združená hustota pozorovaného náhodného výberu*

$$L(\mathbf{x}; \theta) = \prod_{t=1}^T f(x_t; \theta)$$

sa nazýva vierohodnostná funkcia. Potom maximálne vierohodný odhad je hodnota parametru, ktorá maximalizuje vierohodnostnú funkciu.

Definícia 2.10 *Klasický normálny lineárny regresný model spĺňa predpoklady ako klasický lineárny regresný model z definície 2.1, ale navyše, chybový člen má normálne rozdelenie, t.j. $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.*

Veta 2.6 *Nech sú splnené predpoklady klasického normálneho lineárneho regresného modelu. Maximálne vierohodné odhady parametrov β a σ^2 , $\tilde{\beta}$ a $\tilde{\sigma}^2$, majú nasledujúce vlastnosti:*

1. $\tilde{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, kde $\tilde{\beta} = \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,
2. $T\tilde{\sigma}/\sigma^2 = (T - K)\hat{\sigma}^2/\sigma^2 \sim \chi_{T-K}^2$, kde $\tilde{\sigma}^2 = (T - K)\hat{\sigma}^2/T$
3. $\tilde{\beta}$ a $\tilde{\sigma}^2$ sú vzájomne nezávislé

Dôkaz vid' [7].

Distribučné vlastnosti z vety 2.6 poskytujú základ pre odvodenie výberových rozdelení štatistík používaných pri testovaní všeobecných lineárnych hypotéz v klasickom lineárnom regresnom modeli.

2.3.2 Odhady na malých výberoch

Za neprítomnosti nejakého predpokladu o parametrickej forme rozdelenia chybovej zložky \mathbf{e} , (napr. normálne rozdelenie, rovnomerné rozdelenie, ...), derivácia triedy eficientných nestranných odhadov pomocou Gaussovej-Markovovej vety vyžaduje linearitu odhadu, t.j. $\beta^* = \mathbf{H}\mathbf{y}$. Tento striktný predpoklad linearity v istom zmysle kompenzuje nedostatok vedomostí o konkrétnej distribúcii chybovej zložky. Za prítomnosti viacerých explicitných informácií však môže byť predpoklad linearity vynechaný.

Existujú dve rôzne metódy pre zistenie, či je nestranný odhad nestranným odhadom s minimálnym rozptylom (medzi lineárnymi i nelineárnymi odhadmi). V tomto kontexte schválne nepoužívame pojem „najlepší odhad“.

Metóda, ktorá nás bude zaujímať v prvom rade sa nazýva Cramérova-Raova dolná medza. Druhou metódou je metóda úplnosti a suficientnosti (dostatočnosti), ktorú stručne tiež popíšeme, aby sme mohli obe metódy porovnať.

Nestranný odhad s minimálnym rozptylom pomocou Cramérovej-Raovej dolnej medze

Tento prístup je priamočiary a je dvojkrokový. Po prvé, hustota vyšetrovanej náhodnej veličiny musí byť *regulárna funkcia*. Po druhé, ak je hustota regulárna, dolná medza rozptylu je presne určená a budeme ju nazývať *Cramérova-Raova dolná medza*. Ak sa rozptyl nestranného odhadu rovná Cramérovej-Raovej dolnej medzi, jedná sa o nestranný odhad s minimálnym rozptylom. V opačnom prípade účinnosť odhadu nie je determinovaná.

Definícia 2.11 *Hustota $f(x_t; \theta)$ sa nazýva regulárna funkcia práve vtedy, keď platia nasledujúce podmienky:*

- (i) rozsah náhodnej premennej x_t je nezávislý na parametrickom vektore θ
- (ii) hustota $f(x_t; \theta)$ je podľa θ diferencovateľná do 2. rádu a tieto derivácie sú ohraničené nejakou integrovateľnou funkciou \mathbf{x} .

T.j. derivácie $\partial f(x_t; \theta)/\partial \theta_i$, $\partial^2 f(x_t; \theta)/\partial \theta_i \partial \theta_j$ a $\partial f_t^2(\theta)/\partial \theta_i^2$ existujú pre každé i, j a existuje nejaká konštanta A , a tiež existuje nejaká funkcia $M_i(\mathbf{x})$, také, že platí $|\partial f(x_t; \theta)/\partial \theta_i| \leq M_i(\mathbf{x})$.

Veta 2.7 Cramérova-Raova veta. *Nech x_1, x_2, \dots, x_T je náhodný výber s regulárnou hustotou $f(x_t; \theta)$ a s vierohodnostnou funkciou*

$$L(\mathbf{x}; \theta) = \prod_{t=1}^T f(x_t; \theta).$$

Definujme informačnú maticu pre náhodný výber $\mathbf{x} = (x_1, \dots, x_T)$:

$$\mathbf{I}(\theta) = -\mathbb{E} \left[\frac{\partial^2 \ln L(\mathbf{x}; \theta)}{\partial \theta \partial \theta'} \right].$$

Nech $\hat{\theta}$ je nestranný (nevychýlený) odhad θ s kovariančnou maticou Σ . Potom matica $\Sigma - [\mathbf{I}(\theta)]^{-1}$ je pozitívne semidefinítaná. A matica

$$[\mathbf{I}(\theta)]^{-1}$$

je Cramérova-Raova matica dolných medzí.

Dôkaz vid' [7].

Cramérova-Raova veta je niekdy nazývaná Cramérova-Raova nerovnosť, nakoľko stanovuje dolnú medzu pre rozptyl odhadu.

Lemma Mnohorozmerná normálna hustota \mathbf{y} je regulárna vzhľadom k jej prvým deriváciám.

Veta 2.8 Raova-Cramérova dolná medza pre rozptyly $\hat{\beta}$ a $\hat{\sigma}^2$ je $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ a $2\sigma^4/T$.

Dôkaz Výpočtom parciálnych derivácií druhého rádu aplikovaných na vierohodnostnú funkciu $\ln L$ dostaneme informačnú maticu

$$\mathbf{I}(\beta, \sigma^2) = \begin{pmatrix} \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0}' & \frac{T}{2\sigma^4} \end{pmatrix}.$$

Takže Raova-Cramérova dolná medza pre nestranné odhady β a σ^2 je $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ a $2\sigma^4/T$.

Q.E.D.

Z Raovej-Cramérovej vety vyplýva, že $\hat{\beta}$, najlepší nestranný odhad β ako Raova-Cramérova dolná medza existuje a jedná sa o odhad získaný pomocou *OLS* metódy. Naopak, $\text{var } \hat{\sigma}^2 = 2\sigma^4/(T - K) > 2\sigma^4/T$, a teda dolná medza pre nestranný odhad σ^2 nie je nikdy dosiahnutá. V tomto prípade Raova-Cramérova veta neposkytuje verifikáciu účinnosti odhadu $\hat{\sigma}^2$. Takže pre overenie eficientnosti odhadu musíme siahnuť po alternatívnom prístupe: úplnej, suficientnej štatistike. Týmto spôsobom ukážeme, že $\hat{\sigma}^2$ je skutočne eficientný nestranný odhad s najmenším rozptylom.

Nestranný odhad s minimálnym rozptylom pomocou prístupu úplnej-suficientnej štatistiky

Týmto prístupom sa dá ukázať, že aj *OLS* odhad $\hat{\sigma}^2$ parametru σ^2 v normálnom klasickom lineárnom regresnom modeli je eficientný nestranný s minimálnym rozptylom.

Výhodou Cramérovho a Raovho prístupu je ľahšia pochopiteľnosť a aplikovateľnosť. Zatiaľ čo Cramér a Rao vyžaduje regularitu hustoty a aplikáciu Cramérovej-Raovej vety, prístup cez úplnu-suficientnú štatistiku vyžaduje definície úplnosti a suficientnosti a tri vety (Fisherova-Neymanova faktorizačná veta, Raova-Blackwellova veta a Lehmannova-Sheffého veta). Tento prístup je síce zložitejší, ale na druhej strane nedáva žiadne nedeterminované výstupy. Vždy vieme určiť, či sa jedná o nestranný odhad s najmenším rozptylom alebo nie. Obvykle je teda tento prístup silnejší.

2.3.3 Testovanie hypotéz

V predchádzajúcej časti sme sa zameriavali na odhady na malých výberoch, pričom chyby v lineárnom modeli mali normálne rozdelenie. Distribúcie *OLS* odhadov $\hat{\beta}$ a $\hat{\sigma}^2$ v klasickom normálnom lineárnom regresnom modeli boli stanovené pre nejaký konečný rozsah výberu T a tieto odhady boli nestranné a s minimálnym rozptylom. Niekedy máme k dispozícii nejaké dodatočné informácie o konštruovanom modeli. Takéto informácie majú zvyčajne tvar apriórnych obmedzení pre parametre modelu. Najjednoduchším prípadom takýchto obmedzení sú obmedzenia v tvare lineárnych rovností. V praxi je pomerne častá potreba obmedzenia takéhoto typu štatisticky testovať.

V tejto časti uvedieme procedúry vhodné na testovanie takýchto všeobecných hypotéz linearity v klasickom lineárnom modeli. Hypotézu linearity vektorového parametru β môžeme zapísať v tvare $R\beta = r$, kde R je matica konštánt typu $J \times K$ s hodnotou J ($J \leq K$) a r je vektor konštánt typu $J \times 1$. Maticový zápis $R\beta = r$ reprezentuje J nezávislých lineárnych hypotéz o parametri β . Konkrétnym príkladom hypotézy linearity môže byť napríklad, hypotéza, že jeden konkrétny element vektoru r , napr. k -tý prvok, je rovný nejakej konkrétnej konštante, povedzme β_k^* , a ostatné prvky sú nulové (t.j. $R = [0 \ 0 \ \dots \ 0 \ 1 \ 0 \ \dots \ 0]$ a $r = [\beta_k^*]$). V tomto prípade teda $H_0 : \beta_k = R\beta = r = \beta_k^*$. Túto hypotézu H_0 testujeme proti $H_1 : R\beta \neq r$.

Ďalším príkladom lineárnej hypotézy je $H_0 : \beta_i = \beta_i^*, i = 1, 2, \dots, J$; kde $J \leq K$. V tomto prípade alternatíva $H_1 : R\beta \neq r$ proti ktorej testujeme H_0 , znamená, že aspoň jedna z podmienok v H_0 neplatí. Takéto hypotézy, ktoré testujú významnosť niektorých parametrov z β sú veľmi časté. Tieto hypotézy sa dajú testovať F -testom, v praxi však bývajú častejšie používané testy založené na princípe maximálnej vierohodnosti, ktoré fungujú bez ohľadu na tvar modelu. Metóda maximálnej vierohodnosti poskytuje strednú hodnotu získaných bodových odhadov, kým metóda pomeru vierohodností ponúka strednú hodnotu získaných testových štatistík. Pomerom vierohodností sa myslí podiel

$$l = \frac{L(\hat{\omega})}{L(\hat{\Omega})};$$

kde $L(\hat{\omega})$ je maximálna hodnota vierohodnostnej funkcie nad zúženou množinou parametrov, ktorá je definovaná ako tá časť parametrického priestoru, kde platí nulová hypotéza $H_0 : R\beta = r$ a $L(\hat{\Omega})$ je maximálna hodnota vierohodnostnej funkcie na celom parametrickom priestore. Takže $0 \leq l \leq 1$ a l je obstojnou testovou štatistikou pre H_0 . Väčšie hodnoty l znamenajú, že H_0 je zmysluplne prijateľná. Kritickou oblasťou testovej štatistiky je potom $l \leq c_\alpha$, kde c_α určuje z rozdelenia $g(l)$ hladinu testu α , t.j.

$$\int_0^{c_\alpha} g(l) dl = \alpha. \quad (2.2)$$

Túto metódu môžeme aplikovať na všeobecnú lineárnu hypotézu $R\beta = r$ o parametroch klasického normálneho lineárneho modelu $y = X\beta + e$, kde e má mnohorozmerné normálne rozdelenie s nulovou strednou hodnotou a kovariančnou maticou

$\sigma^2 \mathbf{I}$. Všeobecná lineární hypotéza je

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r} \quad (2.3)$$

a alternativa k nej je

$$H_0 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}. \quad (2.4)$$

Vierohodnostná funkcia pre tento model je potom

$$L = (2\pi)^{-T/2} (\sigma^2)^{-T/2} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right]. \quad (2.5)$$

Hodnoty parametrov $\boldsymbol{\beta}$ a σ^2 , ktoré maximalizujú vierohodnostnú funkciu na celom parametrickom priestore sú:

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}, \\ \tilde{\sigma}^2 &= \frac{1}{T} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \end{aligned} \quad (2.6)$$

Substitúciou (2.6) do vierohodnostnej funkcie (2.5) dostaneme

$$L(\hat{\boldsymbol{\Omega}}) = (2\pi)^{-T/2} \left[\frac{1}{T} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \right]^{-T/2} \exp -T/2. \quad (2.7)$$

Hodnoty parametrov $\boldsymbol{\beta}$ a σ^2 , ktoré maximalizujú vierohodnostnú funkciu na parametrickom priestore obmedzenom podmienkou $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ sú

$$\begin{aligned} \boldsymbol{\beta}^* &= \hat{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R} [\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\beta}}), \\ \sigma^{2*} &= \frac{1}{T} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*). \end{aligned} \quad (2.8)$$

Substitúciou týchto odhadov späť do vierohodnostnej funkcie (2.5) dostaneme

$$L(\hat{\omega}) = (2\pi)^{-T/2} \left[\frac{1}{T} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*) \right]^{-T/2} \exp -T/2. \quad (2.9)$$

Potom testová štatistika založená na vierohodnostnom pomere je

$$l = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left[\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})} \right]^{-T/2}.$$

Test založený na pomere vierohodnosti teda zamietá H_0 na hladine α , keď $l \leq c_\alpha$, alebo keď

$$l^{-2/T} = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})} \geq c_\alpha^{-2/T}. \quad (2.10)$$

Pretože

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)' \mathbf{X}'\mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$$

túto štatistiku môžeme upraviť do vhodnejšej formy:

$$I^{-2/T} - 1 = \frac{(\hat{\beta} - \beta^*)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta^*)}{(\mathbf{y} - \mathbf{X} \hat{\beta})' (\mathbf{y} - \mathbf{X} \hat{\beta})} \geq c_{\alpha}^{-2/T} - 1 = c_{\alpha}^*.$$

Ďalej použijeme substitúciu za $(\hat{\beta} - \beta^*)$ z (2.8) a v čitateli dostaneme

$$(\mathbf{r} - \mathbf{R} \hat{\beta})' [\mathbf{R} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{r} - \mathbf{R} \hat{\beta}),$$

a teda test založený na podiele maximálnej vierohodnosti zamietajú H_0 ak

$$\frac{(\mathbf{r} - \mathbf{R} \hat{\beta})' [\mathbf{R} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{r} - \mathbf{R} \hat{\beta}) / J}{(\mathbf{y} - \mathbf{X} \hat{\beta})' (\mathbf{y} - \mathbf{X} \hat{\beta}) / (T - K)} \geq c_{\alpha}^{**}. \quad (2.11)$$

K odvodeniu rozdelenia testovej štatistiky z (2.11) potrebujeme nasledujúce dve lemy:

Lemma Nech $\mathbf{x} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_T)$ a \mathbf{A} a \mathbf{B} sú symetrické matice. Potom $Q_1 = \mathbf{x}' \mathbf{A} \mathbf{x} / \sigma^2$ a $Q_2 = \mathbf{x}' \mathbf{B} \mathbf{x} / \sigma^2$ sú nezávislé práve vtedy, keď $\mathbf{A} \mathbf{B} = \mathbf{0}$, kde $\mathbf{0}$ je nulová matica.

Lemma Ak \mathbf{M} je symetrická idempotentná matica rádu $T \times T$ a \mathbf{x} má normálne rozdelenie $N(\mathbf{0}, \sigma^2 \mathbf{I})$, potom kvadratická forma

$$\frac{\mathbf{x}' \mathbf{M} \mathbf{x}}{\sigma^2}$$

má rozdelenie χ^2 o $m = \text{tr} \mathbf{M}$ stupňoch voľnosti.

Teraz môžeme jednoducho odvodiť rozdelenie testovej štatistiky. Ak $\mathbf{r} = \mathbf{R} \beta$, $\mathbf{r} - \mathbf{R} \hat{\beta} = \mathbf{R}(\beta - \hat{\beta}) = -\mathbf{R}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{e}$ a kvadratickú formu z čitateľa (2.11) môžeme zapísať v tvare $\mathbf{e}' \mathbf{A} \mathbf{e} / J$, kde

$$\mathbf{A} = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{R}' [\mathbf{R} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R}]^{-1} \mathbf{R} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'.$$

Matica \mathbf{A} je zjavne symetrická a idempotentná, navyše $\mathbf{e}' \mathbf{A} \mathbf{e} / J$ má χ^2 rozdelenie o $\text{tr} \mathbf{A} = J$ stupňoch voľnosti. Podobne kvadratickú formu z menovateľa (2.11) môžeme označiť ako $\mathbf{e}' \mathbf{B} \mathbf{e} / (T - K)$, kde $\mathbf{B} = \mathbf{I} - \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$ je symetrická a idempotentná matica s hodnotou $(T - K)$, takže rozdelenie $\mathbf{e}' \mathbf{B} \mathbf{e}$ je χ^2_{T-K} . Keďže \mathbf{A} a \mathbf{B} sú idempotentné, $\mathbf{A} \mathbf{B} = \mathbf{0}$. Takže za platnosti nulovej hypotézy, testová štatistika

$$u = \frac{(\mathbf{r} - \mathbf{R} \hat{\beta})' [\mathbf{R} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{r} - \mathbf{R} \hat{\beta}) / J}{(\mathbf{y} - \mathbf{X} \hat{\beta})' (\mathbf{y} - \mathbf{X} \hat{\beta}) / (T - K)}$$

má centrované F -rozdelenie o J a $T - K$ stupňoch voľnosti, pretože je podielom dvoch nezávislých náhodných veličín s χ^2 rozdelením, ktoré sú podelené zodpovedajúcimi stupňami voľnosti. Z toho vyplýva, že nulovú hypotézu zamietame na základe testu založeného na podiele vierohodností na hladine α s kritickou hodnotou c_{α}^{**} .

2.3.4 Odhady a testovanie hypotéz na veľkých výberoch

V tejto časti uvedieme základné definície a tvrdenia týkajúce sa teórie asymptotického rozdelenia, asymptotických vlastností maximálne vierohodných odhadov a asymptotickej efektivity OLS odhadov parametrov β a σ^2 za predpokladu normality.

Definícia 2.12 *Nech (x_1, x_2, \dots) je postupnosť náhodných veličín s kumulatívnymi distribučnými funkciami (F_1, F_2, \dots) . Potom postupnosť (x_1, x_2, \dots) konverguje v distribúcii ku kumulatívnej distribučnej funkcii F , ak*

$$\lim_{t \rightarrow \infty} F_t = F$$

pre všetky body spojitosti F .

Zjednodušene táto definícia znamená, že kumulatívne distribučné funkcie náhodných veličín x_1, x_2, \dots konvergujú ku kumulatívnej distribučnej funkcii F nejakej náhodnej veličiny x , a teda hovoríme, že náhodné veličiny x_1, x_2, \dots konvergujú v distribúcii k náhodnej veličine x .

Definícia 2.13 *Nech $\hat{\theta}_T$ je konzistentný odhad θ a nech ďalej $\sqrt{T}(\hat{\theta}_T - \theta)$ konverguje v distribúcii k $N(0, \Sigma)$. Potom hovoríme, že $\hat{\theta}_T$ má asymptotické rozdelenie $N(\theta, \Sigma/T)$.*

Definícia 2.14 *Odhad $\hat{\theta}_T$ nazývame konzistentný, rovnomerný, asymptoticky normálny odhad, keď*

- (i) *je konzistentný,*
- (ii) *$\sqrt{T}(\hat{\theta}_T - \theta)$ konverguje v distribúcii k $N(0, \Sigma)$ a*
- (iii) *ak konverguje rovnomerne v každej kompaktnej podmnožine parametrického priestoru.*

Definícia 2.15 *Nech $\hat{\theta}_T$ a $\tilde{\theta}_T$ sú konzistentné, rovnomerné, asymptoticky normálne odhady s asymptotickými kovariančnými maticami Σ/T a Ω/T . Potom hovoríme, že $\hat{\theta}_T$ je asymptoticky efektívny vzhľadom k $\tilde{\theta}_T$, ak matica $\Omega - \Sigma$ je pozitívne definitná.*

Definícia 2.16 *Konzistentný, rovnomerný, asymptoticky normálny odhad nazývame asymptoticky efektívny, ak je asymptoticky efektívny vzhľadom k nejakému inému konzistentnému, rovnomerne asymptoticky normálnemu odhadu.*

Dôsledok Raovej–Cramérovej vety Postačujúcou podmienkou aby konzistentný, rovnomerný, asymptoticky normálny odhad bol asymptoticky efektívnym odhadom je rovnosť asymptotickej kovariancie a Cramérovej–Raovej dolnej medze

$$1/T \lim_{T \rightarrow \infty} [\mathbf{I}(\theta)/T]^{-1}.$$

Nasledujúca veta ukazuje, akým účinným nástrojom je maximálne vierohodný odhad.

Veta 2.9 *Nech x_1, x_2, \dots, x_T je náhodný výber z rozdelenia $f(x_t; \boldsymbol{\theta})$. Nech ďalej $\hat{\boldsymbol{\theta}}_T$ je maximálne vierohodný odhad parametru $\boldsymbol{\theta}$. Potom $\hat{\boldsymbol{\theta}}_T$ je konzistentný odhad $\boldsymbol{\theta}$ a $\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta})$ má asymptotické rozdelenie $N(\mathbf{0}, \lim_{T \rightarrow \infty} [\mathbf{I}(\boldsymbol{\theta})/T]^{-1})$ ak sú splnené „podmienky regularity“.*

Maximálne vierohodné podmienky regularity sú:

- (i) doména náhodnej veličiny x_t nezávisí na vektorovom parametre $\boldsymbol{\theta}$,
- (ii) hustota $f(x_t; \boldsymbol{\theta})$ má derivácie do tretieho rádu podľa $\boldsymbol{\theta}$, ktoré sú ohraňované nejakou integrovateľnou funkciou \mathbf{x} .

Tieto podmienky regularity sa líšia od podmienok z Raovej–Cramérovej vety iba v existencii derivácií tretieho rádu, ktoré sú ohraňované nejakou integrovateľnou funkciou x .

Veta 2.10 Slutského veta. *Ak $\hat{\theta}$ je konzistentný odhad parametru θ , potom $g(\hat{\theta})$ je konzistentným odhadom $g(\theta)$, kde g je spojitá funkcia.*

Slutského veta hovorí, že konzistencia odhadu je invariantná vzhľadom ku spojitaj transformácii. Túto vetu môžeme aplikovať aj na vektorové parametre. Z nasledujúcej vety zasa uvidíme, že funkcia maximálne vierohodných odhadov je maximálne vierohodný odhad funkcie.

Veta 2.11 Invariantnosť maximálne vierohodného odhadu.

Nech $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_K$ sú maximálne vierohodné odhady parametrov $\theta_1, \theta_2, \dots, \theta_K$. Nech ďalej $\alpha_1 = \alpha_1(\theta_1, \theta_2, \dots, \theta_K), \dots, \alpha_K = \alpha_K(\theta_1, \theta_2, \dots, \theta_K)$ je množina transformácií. Potom maximálne vierohodné odhady $\alpha_1, \dots, \alpha_K$ sú $\hat{\alpha}_1 = \alpha_1(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_K), \dots, \hat{\alpha}_K = \alpha_K(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_K)$.

Dôsledok vety 2.4. Nech $\hat{\boldsymbol{\theta}}$ je maximálne vierohodný odhad parametru $\boldsymbol{\theta}$. Potom asymptotická kovariančná matica vektorovej funkcie $\boldsymbol{\psi}(\hat{\boldsymbol{\theta}})$ takej, že $\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}) = (\psi_1(\hat{\boldsymbol{\theta}}), \dots, \psi_K(\hat{\boldsymbol{\theta}}))'$ a $\psi_i(\hat{\boldsymbol{\theta}})$, $i = 1, 2, \dots, K$ sú vzájomne nezávislé skalárne funkcie $\boldsymbol{\theta}$, je

$$\frac{1}{T} \mathbf{G} \lim_{T \rightarrow \infty} \left[-\frac{1}{T} \mathbf{E} \frac{\partial^2 \ln L(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]^{-1} \mathbf{G}',$$

kde (i, j) -tý prvok, $i, j = 1, 2, \dots, K$ matice \mathbf{G} označuje deriváciu $\psi_i(\boldsymbol{\theta})$ podľa $\boldsymbol{\theta}$. Navyše $\boldsymbol{\psi}(\hat{\boldsymbol{\theta}})$ je asymptoticky eficientný odhad $\boldsymbol{\psi}(\boldsymbol{\theta})$.

Ďalšia veta zovšeobecňuje vlastnosti štatistiky založenej na podiele vierohodností na malých výberoch.

Veta 2.12 *Predpokladajme, že platia podmienky regularity maximálne vierohodného odhadu. Nech $l = L(\hat{\omega})/L(\hat{\Omega})$ označuje štatistiku založenú na podiele vierohodností,*

kde Ω je všeobecná K –rozmerná parametrická množina a množina ω je $(K - q)$ –rozmerná podmnožina Ω zodpovedajúca platnosti nulovej hypotézy. Potom, za platnosti nulovej hypotézy asymptoticky platí:

$$-2 \ln l \sim \chi_q^2.$$

Z predchádzajúceho textu je zrejmé, že sú dva spôsoby, ktorými môžeme ukázať, že konzistentný, rovnomerný, asymptoticky normálny odhad je asymptoticky eficientný. Buď dokážeme, že skúmaný odhad má rovnaké asymptotické rozdelenie ako maximálne vierohodný odhad alebo ukážeme, že rozptyl odhadu dosahuje Cramérovu–Raovu dolnú medzu. Oba prístupy môžeme použiť pri skúmaní asymptotickej eficientnosti OLS odhadov $\hat{\beta}$ a $\hat{\sigma}^2$ v klasickom normálnom lineárnom modeli (teda za splnenia predpokladov o chybovej zložke regresného modelu).

Pretože $\hat{\beta}$ a $\hat{\sigma}^2 = (\mathbf{y} - X\hat{\beta})'(\mathbf{y} - X\hat{\beta})/T$ sú maximálne vierohodné odhady, asymptotická eficientnosť vyplýva priamo z vety 2.4 a zo skutočnosti, že viacrozmerné normálne rozdelenie spĺňa podmienky regularity. Nakoľko $\hat{\sigma}^2$ limitne konverguje k maximálne vierohodnému odhadu $\tilde{\sigma}^2$, odhad $\hat{\sigma}^2$ je tiež asymptoticky eficientný.

Prípadne môžeme k overeniu asymptotickej eficientnosti odhadov $\hat{\beta}$ a $\hat{\sigma}^2$ použiť Cramérovu–Raovu dolnú medzu. Informačná matica je

$$\mathbf{I}(\beta, \sigma^2) = \begin{pmatrix} \frac{X'X}{\sigma^2} & \mathbf{0} \\ \mathbf{0}' & \frac{T}{2\sigma^4} \end{pmatrix},$$

teda

$$\lim_{T \rightarrow \infty} \left[\frac{\mathbf{I}(\beta, \sigma^2)}{T} \right] = \begin{pmatrix} \frac{Q}{\sigma^2} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma^4} \end{pmatrix}$$

a asymptotická Cramérovu–Raovu dolná medza je

$$\frac{1}{T} \lim_{T \rightarrow \infty} \left[\frac{\mathbf{I}(\beta, \sigma^2)}{T} \right] = \frac{1}{T} \begin{pmatrix} \sigma^2 \mathbf{Q}^{-1} & \mathbf{0} \\ \mathbf{0}' & 2\sigma^4 \end{pmatrix}.$$

Veta 2.13 *Asymptotické rozdelenie $\sqrt{T}(\hat{\beta} - \beta)$ je centrované normálne s konečným rozptylom $\sigma^2 \mathbf{Q}^{-1}$, t.j.*

$$\sqrt{T}(\hat{\beta} - \beta) \sim N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}),$$

kde $\mathbf{Q} = \lim_{T \rightarrow \infty} (X'X/T)$ a $\hat{\beta}$ je asymptoticky eficientný odhad parametru β .

Lemma *Lindebergova-Lévyho centrálna limitná veta.* Nech x_1, x_2, \dots je postupnosť i.i.d. náhodných veličín s konečnou strednou hodnotou a s konečným rozptylom σ^2 . Potom náhodná veličina

$$z_t = \frac{1}{\sqrt{T}} \sum_{t=1}^T \left(\frac{x_t - \mu}{\sigma} \right)$$

konverguje v distribúcii k $N(0, 1)$.

Veta 2.14 *Asymptotické rozdelenie $\sqrt{T}(\hat{\sigma}^2 - \sigma^2)$ je $N(0, 2\sigma^4)$ a $\hat{\sigma}^2$ je asymptoticky eficientný odhad σ^2 .*

Asymptotická účinnosť $\hat{\beta}$ a $\hat{\sigma}^2$ závisí na ich koincidencii s maximálne virohodnými odhadmi. Tieto výsledky sú úplne závislé na predpoklade normality. Napríklad, ak je daný štandardný lineárny model s homoskedastickými chybami ale s exponenciálnym rozdelením, asymptotická eficientnosť sa nezachová.

Vidíme, že normalita je základným prvkom pri testovaní hypotéz a vlastnostiach odhadov na malých výberoch. Vzniká teda otázka, aké vlastnosti lineárneho modelu sú zachované pokiaľ vynecháme predpoklad normality.

Z predchádzajúcich častí už vieme, že $\hat{\beta}$ je najlepší nestranný lineárny a konzistentný odhad parametru β a má kovariančnú maticu $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, zatiaľ čo $\hat{\sigma}^2$ je nestranný a konzistentný odhad. Avšak $\hat{\beta}$ nemá normálne rozdelenie, ani $(T - K)\hat{\sigma}^2/\sigma^2$ nemá χ^2 rozdelenie a ani $\hat{\beta}$ a $\hat{\sigma}^2$ nie sú nestranné s minimálnym rozptylom a ani asymptoticky eficientné. Navyše nebudú tu ďalej platiť ani testy vyvinuté za predpokladu normality na malých výberoch. Zvlášť, ak poznáme nenormálne rozdelenie chybovej zložky \mathbf{e} , potom môžeme stanoviť výberovú distribúciu pre $\hat{\beta}$ a $\hat{\sigma}^2$ a teda aj konfidenčné intervaly. Či to dokážeme alebo nedokážeme stanoviť závisí na priechodnosti požadovaného rozdelenia. Navyše uvedené procedúry zostávajú aspoň asymptoticky platné za podmienky, že chyby sú *i.i.d.* s nulovou strednou hodnotou a konštantným rozptylom σ^2 . Takže potrebujeme zabezpečiť homoskedasticitu chybovej zložky.

Kapitola 3

Porušenie základných predpokladov

Klasický lineárny regresný model a klasický normálny lineárny regresný model popisujú klasické predpoklady regresnej analýzy. V klasickom lineárnom modeli sme predpokladali, že všetky chyby majú rovnaký rozptyl, sú nekorelované (resp. nezávislé) a majú normálne rozdelenie s nulovou strednou hodnotou. Oba modely, zhodne s Aitkenovou vetou, používajú metódu najmenších štvorcov (*OLS*). Pri použití *OLS* v prítomnosti nesférických chybových vplyvov klasického regresného modelu odhad stráca účinnosť a model neposkytuje vhodný základ pre štatistické závery. Dva základné charakteristické typy nesférických chýb sú heteroskedasticita a autokorelácia. Jedna z metód odhadovania používaná za prítomnosti heteroskedasticity a autokorelácie sa nazýva „Feasible“ (funkčná, použiteľná) zobecnená metóda najmenších štvorcov. Pojednáme o nej ďalej. Ďalšími porušeniami predpokladov klasického lineárneho modelu sú súbežné korelácie medzi chybovou zložkou a vysvetľujúcou premennou. Za takýchto okolností je metóda najmenších štvorcov nekonzistentná a neúčinná. Dva prípady, keď takéto korelácie nastanú, sú keď koeficienty musia byť odhadované v omeškanej závislej premennej (model postupných korelácií - serial correlation model) a v modeli nepozorovateľnej premennej. Ďalšou možnosťou ako pristupovať k porušeniu predpokladov kladených na reziduálnu zložku klasického regresného modelu, je zobrať ako základný prvok konštrukcie modelu práve reziduálnu zložku, ktorá môže byť tvorená korelovanými náhodnými veličinami. V tomto prípade sa jedná o Boxovu–Jenkinsovú metodológiu, ktorá je veľmi flexibilná. Porušeniami predpokladov klasického lineárneho modelu a metódami, ktoré sa pri porušení predpokladov používajú sa budeme zaoberať v tejto kapitole.

3.1 „Feasible“ zobecnená metóda najmenších štvorcov

V niektorých prípadoch chceme špecifikovať štruktúru kovariančnej matice chýb, Ω , v zobecnenom modeli najmenších štvorcov (*GLM*) až na nejaké neznáme parametre, $\theta_1, \theta_2, \dots, \theta_p$. Nastáva to napríklad pri podozrení na korelovanosť chýb v regresnom

modele časových radov alebo ak má regresný vzťah premenlivú presnosť. Ďalej sa budeme zaoberať s kovariančnou maticou chýb, Ω , pri všeobecných podmienkach, jej odhadom pokiaľ je Ω neznáma, a konsekvenciami tohoto odhadu s vlastnosťami odhadov parametrov v *GLM*.

3.1.1 Definícia „Feasible“ *GLM*

Pripomeňme, že zobecnený model najmenších štvorcov má tvar $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, kde \mathbf{X} je nenáhodná matica typu $T \times K$ s hodnotou K a platí

$$\lim_{T \rightarrow \infty} \left(\frac{\mathbf{X}'\Omega^{-1}\mathbf{X}}{T} \right) = \mathbf{Q}_\Omega,$$

\mathbf{Q}_Ω je konečná regulárna matica typu $K \times K$. Chybový vektor \mathbf{e} spĺňa podmienky: $\mathbf{E}\mathbf{e} = \mathbf{0}$ a $\mathbf{E}\mathbf{e}\mathbf{e}' = \sigma^2\Omega$. Miesto podmienok, že chyby sú *i.i.d.* s nulovou strednou hodnotou a s konečným rozptylom σ^2 predpokladajme radšej, že chyby e_t pochádzajú z autoregresného procesu 1. rádu $e_t = \rho e_{t-1} + u_t$, $|\rho| < 1$, kde u_t sú *i.i.d.* s nulovou strednou hodnotou a s konečným rozptylom σ_u^2 . V tejto sekcii tiež predpokladáme, že kovariančná matica chýb, Ω , je známa, až na nejaký skalárny násobok (aj napriek tomu, že najčastejšie býva Ω uvažovaná ako neznáma). Formálne:

Definícia 3.1 „Feasible“ zovšeobecnený model najmenších štvorcov (*FGLM*) je definovaný ako

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kde \mathbf{y} je vektor pozorovaní závislej premennej rádu $T \times 1$, $\boldsymbol{\beta}$ je vektor neznámych koeficientov rádu $K \times 1$, \mathbf{X} je nenáhodná matica pozorovaní vysvetľujúcich premenných rádu $T \times K$, s hodnotou K , a konečne \mathbf{e} je autoregresný proces prvého rádu, $e_t = \rho e_{t-1} + u_t$, taký, že $|\rho| < 1$, a u_t sú centrovane *i.i.d.* náhodné veličiny s konečným rozptylom, t.j. $\mathbf{E}u_t = 0$ a $\text{var}u_t = \sigma_u^2 \leq \infty$ a $\mathbf{E}\mathbf{e}\mathbf{e}' = \sigma_u^2\Omega$, kde

$$\sigma_u^2\Omega = \frac{\sigma_u^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & & \vdots \\ \vdots & \vdots & & \ddots & \rho \\ \rho^{T-1} & \rho^{T-2} & \dots & \rho & 1 \end{bmatrix}.$$

Definícia 3.2 Ak Ω závisí na konečnom počte parametrov $\theta_1, \theta_2, \dots, \theta_p$ a ak $\hat{\Omega}$ závisí na konzistentných odhadoch týchto parametrov, $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p$, potom $\hat{\Omega}$ je konzistentný odhad Ω .

Definícia 3.3 Nech $\hat{\Omega}$ je konzistentný odhad Ω . Potom odhad „feasible“ zovšeobecnenou metódou najmenších štvorcov (resp. *FGLS*-odhad) parametru $\boldsymbol{\beta}$ je

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y}.$$

Ak by sme použili bežnú metódu najmenších štvorcov, dostaneme:

$$\hat{\Omega} = (\mathbf{y} - X\hat{\beta})(\mathbf{y} - X\hat{\beta})'.$$

V tomto prípade $h(\hat{\Omega}) = 1$ a preto $\hat{\Omega}$ je singulárna, takže k nej neexistuje inverzná matica a preto nevieme vypočítať ani odhad $\hat{\beta}$. Taktiež máme $T(T+1)/2$ rôznych prvkov v Ω , ale iba T pozorovaní. Takže počet parametrov v neštrukturovanom zobecnenom modeli najmenších štvorcov je $T(T+1)/2 + K > T$ a bez ohľadu na rozsah výberu budeme mať nedostatočný počet stupňov voľnosti.

Zrejme pre Ω musia platiť nasledujúce minimálne obmedzenia:

- (i) $h(\hat{\Omega}) = T$ (takže $\hat{\Omega}$ je regulárna, a teda $\hat{\Omega}^{-1}$ existuje),
- (ii) pre počet p parametrov v matici Ω , musí platiť: $K + p \leq T$ a
- (iii) musia existovať konzistentné odhady parametrov v matici Ω .

3.1.2 Vlastnosti odhadu „feasible“ zovšeobecnenou metódou najmenších štvorcov

V prípade splnenia podmienok pre *GLM*, *GLS*-odhad $\hat{\beta}$ je najlepším nestranným lineárnym odhadom parametru β . Ak pridáme predpoklad, že chyby sú normálne, *GLS*-odhad má ďalšiu vlastnosť - je nestranným odhadom s najmenším rozptylom (medzi lineárnymi i nelineárnymi odhadmi). Ďalej budeme skúmať, ktoré vlastnosti *GLS*-odhadu sú zachované a ktoré stratíme, keď Ω nepoznáme, a teda ju odhadujeme.

***FGLS* za absencie normálnej distribúcie chýb**

Uvažujme *GLM*. Nepotrebuje žiadne špecifické predpoklady na distribúciu chybovej zložky k odvodeniu viacerých použiteľných vlastností „feasible“ zovšeobecnenej metódy najmenších štvorcov. Aitkenovu vetu však nemôžeme použiť, nakoľko Ω je neznáma a musíme ju odhadovať. Všeobecné vlastnosti *FGLM* popisuje nasledujúca veta.

Veta 3.1 „Feasible“ zovšeobecnená metóda najmenších štvorcov pri absencii normálneho rozdelenia chybovej zložky. *Nech platia podmienky zovšeobecneného modelu najmenších štvorcov a nech ďalej platí:*

- (i) Existuje matica \mathbf{P}' taká, že $\mathbf{P}\mathbf{P}' = \Omega^{-1}$ a prvky matice $\mathbf{P}'\mathbf{e}$ sú i.i.d.
- (ii) $\text{plim} \frac{\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}}{T} = \text{plim} \frac{\mathbf{X}'\Omega^{-1}\mathbf{X}}{T} = \mathbf{Q}_\Omega$, kde \mathbf{Q}_Ω je konečná a regulárna
- (iii) $\text{plim} \frac{\mathbf{X}'\hat{\Omega}^{-1}\mathbf{e}}{T} = \text{plim} \frac{\mathbf{X}'\Omega^{-1}\mathbf{e}}{T} = \mathbf{0}$
- (iv) $\text{plim} \frac{1}{T}\mathbf{e}'\hat{\Omega}^{-1}\mathbf{e} = \text{plim} \frac{1}{T}\mathbf{e}'\Omega^{-1}\mathbf{e} = \sigma^2$.

Odhad „feasible“ zobecnenou metódou najmenších štvorcov $\tilde{\beta}$ je konzistentný, $\sqrt{T}(\tilde{\beta} - \beta)$ je asymptoticky $N(\mathbf{0}, \sigma^2 Q_{\Omega}^{-1})$ a

$$\tilde{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\tilde{\beta})' \hat{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\tilde{\beta})}{(T - K)}$$

je konzistentný odhad σ^2 . Z asymptotickej normality štatistika $t = \frac{\tilde{\beta}_s - \beta_s^0}{\sqrt{q_{ss}}}$, kde β_s je s -tý prvok β a q_{ss} je s -tý diagonálny prvok $\tilde{\sigma}^2(\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}$, je za nulovej hypotézy, $H_0 : \beta_s = \beta_s^0$, asymptoticky $N(0, 1)$. Pretože t -rozdelenie konverguje k $N(0, 1)$, keď $T \rightarrow \infty$, dostaneme konvenčný t -test signifikantnosti koeficientu. Podobne, testom všeobecnej lineárnej hypotézy $\mathbf{R}\beta = \mathbf{r}$ tiež dostaneme, že štatistika

$$\frac{(\mathbf{R}\tilde{\beta} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\tilde{\beta} - \mathbf{r}) / J}{(\mathbf{y} - \mathbf{X}\tilde{\beta})' \hat{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\tilde{\beta}) / (T - K)}$$

má asymptoticky F -rozdelenie o J a $T - K$ stupňoch voľnosti za nulovej hypotézy $H_0 : \mathbf{R}\beta = \mathbf{r}$.

Dôkaz vety 3.1 je uvedený v [7]. Asymptotické vlastnosti $FGLS$ -odhadu sú teda rovnaké ako u GLS -odhadu. Pri veľkých výberoch, teda môžeme použiť konvenčné testovanie hypotéz. Keď Ω nie je známa, vlastnosť najlepšieho nestranného lineárneho odhadu a aj nestranného odhadu s minimálnym rozptylom strácame, nakoľko sa jedná o vlastnosti odhadov na malých výberoch.

***FGLS* za prítomnosti normálnej distribúcie chýb**

Aj naďalej budeme uvažovať platnosť predpokladov z vety 3.1 a naviac pribudne predpoklad, že chyby \mathbf{e} majú mnohorozmerné normálne rozdelenie. Kompletnú diskusiu o maximálne vierohodnom odhade v GLM s neznámou kovariančnou maticou chýb bez straty všeobecnosti poskytol Magnus, ktorý predpokladal, že $\sigma^2 = 1$ a $\mathbf{E}\mathbf{e}\mathbf{e}' = \Omega$, kde Ω je pozitívne definitná matica s prvkami, ktoré sú dvakrát diferencovateľné funkcie $\theta_1, \theta_2, \dots, \theta_p$.

Veta 3.2 Magnus („Feasible“ zovšeobecnená metóda najmenších štvorcov s normálnymi chybami) Za platnosti predpokladov uvedených vo vete 3.1 a ak chyby \mathbf{e} majú mnohorozmerné normálne rozdelenie, dostaneme riešenie pre maximálne vierohodné podmienky prvého rádu pre β a $\theta' = (\theta_1, \theta_2, \dots, \theta_p)$:

$$\check{\beta} = (\mathbf{X}'\check{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\check{\Omega}^{-1}\mathbf{y}, \quad (3.1)$$

$$tr \left(\frac{\partial \Omega^{-1}}{\partial \theta_h} \Omega \right)_{\theta=\check{\theta}} = \check{\mathbf{e}}' \left(\frac{\partial \Omega^{-1}}{\partial \theta_h} \right)_{\theta=\check{\theta}} \check{\mathbf{e}}, \text{ kde } h = 1, 2, \dots, p \quad (3.2)$$

a $\check{\mathbf{e}} = \mathbf{y} - \mathbf{X}\check{\beta}$. Rovnosť 3.2 bude v ďalšom uvádzaná ako θ -rovnosť. Ďalej, ak $|\Omega|$ nezávisí na θ_h , h -tá rovnosť v θ -rovnosti sa redukuje na:

$$\check{\mathbf{e}}' \left(\frac{\partial \Omega^{-1}}{\partial \theta_h} \right)_{\theta=\check{\theta}} \check{\mathbf{e}} = 0. \quad (3.3)$$

Dôkaz Vierohodnostná funkcia:

$$\begin{aligned} L &= |2\pi|^{-T/2} |\mathbf{\Omega}|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right], \\ \ln L &= -\frac{T}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{\Omega}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \quad (3.4)$$

Podmienky prvého rádu sú:

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \left[-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right] \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} \left[-\frac{1}{2} \mathbf{y}' \mathbf{\Omega}^{-1} \mathbf{y} + \frac{1}{2} \boldsymbol{\beta}' \mathbf{X}' \mathbf{\Omega}^{-1} \mathbf{y} + \frac{1}{2} \mathbf{y}' \mathbf{\Omega}^{-1} \mathbf{X} \boldsymbol{\beta} - \frac{1}{2} \boldsymbol{\beta}' \mathbf{X}' \mathbf{\Omega}^{-1} \mathbf{X} \boldsymbol{\beta} \right] \\ &= \mathbf{X}' \mathbf{\Omega}^{-1} \mathbf{y} - \mathbf{X}' \mathbf{\Omega}^{-1} \mathbf{X} \boldsymbol{\beta}, \end{aligned} \quad (3.5)$$

$$\begin{aligned} \frac{\partial \ln L}{\partial \theta_h} &= \frac{\partial}{\partial \theta_h} \left[-\frac{1}{2} \ln |\mathbf{\Omega}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right] \\ &= \frac{\partial}{\partial \theta_h} \left(\frac{1}{2} \ln |\mathbf{\Omega}^{-1}| \right) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \frac{\partial \mathbf{\Omega}^{-1}}{\partial \theta_h} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{1}{2} \text{tr} \left(\frac{\partial \mathbf{\Omega}^{-1}}{\partial \theta_h} \mathbf{\Omega} \right) - \frac{1}{2} \mathbf{e}' \left(\frac{\partial \mathbf{\Omega}^{-1}}{\partial \theta_h} \right) \mathbf{e}, \quad \text{pre } h = 1, 2, \dots, p \end{aligned} \quad (3.6)$$

Z toho plynie, že maximálne vierohodný odhad $\check{\boldsymbol{\beta}} = (\mathbf{X}' \check{\mathbf{\Omega}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \check{\mathbf{\Omega}}^{-1} \mathbf{y}$ a maximálne vierohodné odhady $\check{\theta}_1, \check{\theta}_2, \dots, \check{\theta}_p$ splňajú p rovníc:

$$\text{tr} \left(\frac{\partial \mathbf{\Omega}^{-1}}{\partial \theta_h} \mathbf{\Omega} \right)_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} = \check{\mathbf{e}}' \left(\frac{\partial \mathbf{\Omega}^{-1}}{\partial \theta_h} \right)_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} \check{\mathbf{e}}, \quad h = 1, 2, \dots, p.$$

Samozrejme, ak $|\mathbf{\Omega}|$ nezávisí na θ_h , platí (3.3)

Q.E.D.

Ak je daná $\mathbf{\Omega}$, normálne rovnosti pre $\check{\boldsymbol{\beta}}$ sú lineárne a preto jednoducho riešiteľné. Avšak θ -rovnosť je výrazne nelineárna a v mnohých prípadoch nejde maximálne vierohodný odhad $\check{\theta}_h$ získať analyticky. V takýchto prípadoch je funkčná nasledujúca iteračná maximálne vierohodná procedúra:

1. Zvoľme $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, množina prípustných hodnôt $\boldsymbol{\theta}$.
2. Počítajme $\mathbf{\Omega}_0^{-1} = \mathbf{\Omega}^{-1}(\boldsymbol{\theta}_0)$

$$\mathbf{b}_0 = (\mathbf{X}' \mathbf{\Omega}_0^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{\Omega}_0^{-1} \mathbf{y},$$

$$\mathbf{e}_0 = \mathbf{y} - \mathbf{X} \mathbf{b}_0.$$

3. Substituujeme \mathbf{e}_0 do θ -rovnosti (3.6). Dostaneme p (nelineárnych) rovností o p neznámych $\theta_1, \theta_2, \dots, \theta_p$. A z toho vypočítame hodnotu vektorového parametru $\boldsymbol{\theta}$, povedzme $\boldsymbol{\theta}_1$.

4. Počítajme $\mathbf{\Omega}_1^{-1} = \mathbf{\Omega}^{-1}(\boldsymbol{\theta}_1)$. $\mathbf{b}_1 = (\mathbf{X}'\mathbf{\Omega}_1^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}_1^{-1}\mathbf{y}$ a pokračujeme, kým algoritmus neskonverguje.

Veta 3.3 Informačná matica vierohodnostnej funkcie

$$\ln L = -\frac{T}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{\Omega}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

je

$$I = \begin{bmatrix} \mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X} & 0 \\ 0 & \frac{1}{2}\boldsymbol{\psi} \end{bmatrix},$$

kde $\boldsymbol{\psi}$ je symetrická matica typu $p \times p$ s prvkami

$$(\boldsymbol{\psi})_{ij} = \text{tr} \left(\frac{\partial \mathbf{\Omega}^{-1}}{\partial \theta_i} \mathbf{\Omega} \frac{\partial \mathbf{\Omega}^{-1}}{\partial \theta_j} \mathbf{\Omega} \right), \quad \text{kde } i, j = 1, 2, \dots, p.$$

Z toho vyplýva, že asymptotická kovariančná matica pre maximálne vierohodné odhady $\check{\boldsymbol{\beta}}$ a $\check{\boldsymbol{\theta}}$ získané iteratívnou procedúrou je

$$\frac{1}{T} \lim_{T \rightarrow \infty} \left(\frac{\mathbf{I}}{T} \right)^{-1}.$$

Konzistentný odhad asymptotickej kovariančnej matice je

$$\check{\mathbf{I}}^{-1} = \begin{bmatrix} (\mathbf{X}'\check{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1} & 0 \\ 0 & 2\check{\boldsymbol{\psi}}^{-1} \end{bmatrix},$$

pričom všetky neznáme parametre sú nahradené svojimi maximálne vierohodnými odhadmi.

Z vety o maximálnej vierohodnosti vyplýva, že $\check{\boldsymbol{\beta}}$ a $\check{\boldsymbol{\theta}}$ sú konzistentné, asymptoticky efektívne a ich rozdelenie je asymptoticky normálne. Potom $\check{\mathbf{I}}^{-1}$ môže byť použitá v spojení s normálnym rozdelením k zostaveniu asymptotických testov.

3.1.3 „Zdanlivo“ nesúvisiace regresie

Skúmame základný ekonometrický model s využitím *FGLS*-odhadu, *zdanlivo* nesúvisiacimi regresiami.

Investičná teória, ktorú uviedol Zellner v roku 1962 predpokladá, že firemné investície v danom roku sú aspoň čiastočne určené očakávaným ziskom v tomto roku;

Pretože očakávaný zisk v ďalšom roku nie je priamo pozorovateľný, nahradíme ho merateľnou tržnou hodnotou na konci roka F_t . Ďalej požadujeme aby základný kapitál v nasledujúcom roku C_{t+1}^* bol lineárne závislý s F_t : $C_{t+1}^* = \alpha_0 + \alpha_1 F_t$. Potom $C_{t+1}^* - C_t = \alpha_0 + \alpha_1 F_t - C_t$, kde C_t označuje základný kapitál v čase t . Predpokladáme, že iba istá časť p_t z požadovanej čistej investície je v danom roku realizovaná. Potom skutočná investícia je $p_1(C_{t+1}^* - C_t) + p_1\alpha_0 + p_1\alpha_1 F_t - p_1C_t$. Navyše časť investícií tvoria pravdepodobne nejaké servisné náklady a pod., v množstve

rovnom nejakej konštantnej časti z reálneho základného kapitálu, $p_2 C_t$. Preto hrubá investícia v nasledujúcom roku je: $I_{t+1} = p_1(C_{t+1}^* - C_t) + p_2 C_t = p_1 \alpha_0 + p_1 \alpha_1 F_t + (p_2 - p_1) C_t$, čo je v stochastickom zápise: $I_t = \beta_1 + \beta_2 F_{t-1} + \beta_3 C_{t-1} + e_t$

Zellner sa zameriaval na povahu chybovej zložky, keď investície firiem sú riadené podobnými okolnosťami (napr. rovnaké odvetvie priemyslu). Uvažujme investičné správanie General Electric a Westinghouse. Označme ich investičné funkcie $I_{ti} = \beta_{1i} + \beta_{2i} F_{t-1,i} + \beta_{3i} C_{t-1,i} + e_{ti}$, kde $i = 1, 2$ označujú investičné funkcie General Electric a Westinghouse v tomto poradí a pre každú firmu máme $t = 1, 2, \dots, T$ pozorovaní. Zaujímavá je kovariančná matica chýb

$$E \mathbf{e}_* \mathbf{e}_*' = E \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} [\mathbf{e}_1' \ \mathbf{e}_2'],$$

kde $\mathbf{e}_i = (e_{1i}, e_{2i}, \dots, e_{Ti})'$, $i = 1, 2$. Jednou z možných foriem takejto kovariančnej matice chýb je

$$E \mathbf{e}_* \mathbf{e}_*' = \begin{bmatrix} \sigma_{11} \mathbf{I}_T & \mathbf{0}_{T \times T} \\ \mathbf{0}_{T \times T} & \sigma_{22} \mathbf{I}_T \end{bmatrix},$$

kde $E e_{ti}^2 = \sigma_{ii}$ pre $t = 1, 2, \dots, T$ a $i = 1, 2$, $E e_{si} e_{ti} = 0$, $s \neq t$, $\mathbf{e}_{si} e_{tj} = 0$ pre $i \neq j$ a pre všetky $s, t = 1, 2, \dots, T$ a $\mathbf{I}_{T \times T}$ je jednotková matica typu $T \times T$. V tomto prípade sú chyby pre každú firmu *i.i.d.* a chyby medzi jednotlivými firmami sú vzájomne nezávislé. Avšak je dosť možné, že existujú náhodné exogénne faktory, ktoré budú ovplyvňovať investície oboch firiem, napr. vládna politika, medzinárodné udalosti, hospodárska aktivita a pod. V tomto prípade platí, že $E e_{ti} e_{tj} = \sigma_{ij}$, $i, j = 1, 2$ a $t = 1, 2, \dots, T$ a skutočná kovariančná matica je

$$\Omega = \begin{bmatrix} \sigma_{11} \mathbf{I}_T & \sigma_{12} \mathbf{I}_T \\ \sigma_{21} \mathbf{I}_T & \sigma_{22} \mathbf{I}_T \end{bmatrix}.$$

Tento fakt ozrejmjuje názov „zdanlivo“ *nesúvisiace regresie*. Pozrime sa teraz na zo-

všeobecnenie zdanlivo nesúvisiacich regresných modelov do M rovníc a definujme *štandardné podmienky pre zdanlivo nesúvisiace regresie*.

Uvažujme M regresných rovností, ktoré popisujú investície M firiem:

$$\begin{aligned} \mathbf{y}_1 &= X_1 \boldsymbol{\beta}_1 + \mathbf{e}_1, \\ \mathbf{y}_2 &= X_2 \boldsymbol{\beta}_2 + \mathbf{e}_2, \\ &\vdots \\ \mathbf{y}_M &= X_M \boldsymbol{\beta}_M + \mathbf{e}_M, \end{aligned} \tag{3.7}$$

kde každé \mathbf{y}_i , pre $i = 1, 2, \dots, M$ je matica typu $T \times 1$, \mathbf{X}_i je typu $T \times K_i$, $\boldsymbol{\beta}_i$ je typu $K_i \times 1$ a \mathbf{e}_i je vektor o T zložkách. Týchto M rovností môžeme prepísať do

maticového zápisu:

$$\mathbf{y}_* = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix}, \quad \mathbf{X}_* = \begin{bmatrix} \mathbf{X}_1 & & & \mathbf{0} \\ & \mathbf{X}_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{X}_M \end{bmatrix}, \quad \boldsymbol{\beta}_* = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix}, \quad \mathbf{e}_* = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_M \end{bmatrix},$$

a

$$\mathbf{y}_* = \mathbf{X}_* \boldsymbol{\beta}_* + \mathbf{e}_* \quad (3.8)$$

Definícia 3.4 Predpokladajme, že zdanlivo nesúvisiaca regresia (3.8) spĺňa nasledujúce podmienky:

- (i) $E(\mathbf{e}_*) = \mathbf{0}$.
- (ii) $E(\mathbf{e}_* \mathbf{e}_*') = \boldsymbol{\Omega}$, kde $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}$ a $\boldsymbol{\Sigma} = [\sigma_{ij}]$, kde $i, j = 1, 2, \dots, M$.
- (iii) Matica \mathbf{X}_* je nenáhodná, $(\mathbf{X}_*' \boldsymbol{\Omega}^{-1} \mathbf{X}_*)^{-1}$ je regulárna a

$$\lim_{T \rightarrow \infty} \left(\frac{\mathbf{X}_*' \boldsymbol{\Omega}^{-1} \mathbf{X}_*}{T} \right)^{-1} = \mathbf{Q}_*$$

je konečná regulárna matica.

Tieto predpoklady sa nazývajú štandardné podmienky pre zdanlivo nesúvisiace regrese.

Detailnejšie, kovariančná matica \mathbf{e}_* je $E \mathbf{e}_* \mathbf{e}_*' = \boldsymbol{\Omega}$, kde

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I} = \begin{bmatrix} \sigma_{11} \mathbf{I} & \sigma_{12} \mathbf{I} & \dots & \sigma_{1M} \mathbf{I} \\ \sigma_{12} \mathbf{I} & \sigma_{22} \mathbf{I} & \dots & \sigma_{2M} \mathbf{I} \\ \vdots & & \ddots & \vdots \\ \sigma_{1M} \mathbf{I} & \sigma_{2M} \mathbf{I} & \dots & \sigma_{MM} \mathbf{I} \end{bmatrix},$$

a

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1M} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2M} \\ \vdots & & \ddots & \vdots \\ \sigma_{1M} & \sigma_{2M} & \dots & \sigma_{MM} \end{bmatrix}.$$

Symbolom \otimes značíme operátor Kroneckerovho súčinu, ktorý je definovaný pre ľubovoľnú maticu $\mathbf{A} = [a_{ij}]$ dimenzie $m \times n$ a maticu $\mathbf{B} = [b_{ij}]$ dimenzie $p \times q$ ako

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11} \mathbf{B} & a_{12} \mathbf{B} & \dots & a_{1n} \mathbf{B} \\ a_{21} \mathbf{B} & a_{22} \mathbf{B} & \dots & a_{2n} \mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} \mathbf{B} & a_{m2} \mathbf{B} & \dots & a_{mn} \mathbf{B} \end{bmatrix}_{mp \times nq}.$$

Veta 3.4 Za platnosti štandardných podmienok pre zdanlivo nesúvisiace regresie, najlepší nestranný lineárny konzistentný a (asymptoticky) normálne rozdelený odhad β_* je

$$\tilde{\beta}_* = (\mathbf{X}'_* \Omega^{-1} \mathbf{X}_*)^{-1} \mathbf{X}'_* \Omega^{-1} \mathbf{y}_* \quad (3.9)$$

s kovariančnou maticou $(\mathbf{X}'_* \Omega^{-1} \mathbf{X}_*)^{-1}$ a s asymptotickou kovariančnou maticou

$$\frac{1}{T} \lim_{T \rightarrow \infty} \left(\frac{(\mathbf{X}'_* \Omega^{-1} \mathbf{X}_*)^{-1}}{T} \right)^{-1}. \quad (3.10)$$

Ak sú navyše chyby \mathbf{e}_* normálne rozdelené, potom $\tilde{\beta}_*$ je nestranný odhad s minimálnym rozptylom, ktorý je navyše asymptoticky eficientný.

Odhad zobecnenou metódou najmenších štvorcov $\tilde{\beta}_*$ je vo všeobecnosti účinnejší ako odhad bežnou metódou najmenších štvorcov $\hat{\beta}_*$:

$$\begin{aligned} \hat{\beta}_* = (\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{X}'_* \mathbf{y}_* &= \begin{bmatrix} \mathbf{X}'_1 \mathbf{X}_1 & & & \\ & \mathbf{X}'_2 \mathbf{X}_2 & \mathbf{0} & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{X}'_M \mathbf{X}_M \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'_1 \mathbf{y}_1 \\ \mathbf{X}'_2 \mathbf{y}_2 \\ \vdots \\ \mathbf{X}'_M \mathbf{y}_M \end{bmatrix} = \\ &= \begin{bmatrix} (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y}_1 \\ (\mathbf{X}'_2 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{y}_2 \\ \vdots \\ (\mathbf{X}'_M \mathbf{X}_M)^{-1} \mathbf{X}'_M \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_M \end{bmatrix}, \end{aligned} \quad (3.11)$$

kde $\hat{\beta}_i = (\mathbf{X}'_i \mathbf{X}_i)^{-1} \mathbf{X}'_i \mathbf{y}_i$, pre $i = 1, 2, \dots, M$ je OLS-odhad parametru β z i -tej rovnosti.

V dvoch prípadoch sú tieto odhady identické, a teda aj rovnako účinné.

Veta 3.5 Ak $\sigma_{jk} = 0$ pre každé $j \neq k$, potom $\tilde{\beta}_* = \hat{\beta}_*$ a bežná metóda najmenších štvorcov (OLS) je plne eficientná.

Ďalší prípad, kde odhad zobecnenou metódou najmenších štvorcov $\tilde{\beta}_*$ a odhad bežnou metódou najmenších štvorcov sú numericky rovné a teda aj rovnako účinné je, keď regresné matice \mathbf{X}_i , $i = 1, 2, \dots, M$ sú numericky identické, čo je formálne zapísané v nasledujúcej vete.

Veta 3.6 Uvažujme množinu rovností:

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X} \beta_1 + \mathbf{e}_1 \\ \mathbf{y}_2 &= \mathbf{X} \beta_2 + \mathbf{e}_2 \\ &\vdots \\ \mathbf{y}_M &= \mathbf{X} \beta_M + \mathbf{e}_M. \end{aligned}$$

V tomto prípade $\tilde{\beta}_* = \hat{\beta}_*$ a bežná metóda najmenších štvorcov je plne eficientná.

Pokiaľ pre všetkých M konštrukčných matíc z predchádzajúcej definície platí, že sú jednotkové, t.j. $\mathbf{X}_1 = \mathbf{X}_2 = \dots = \mathbf{X}_M = \mathbf{X}$, táto veta platí aj napriek množstvu korelácií v chybovej zložke.

3.1.4 Odhady: asymptotické výsledky

V tejto časti budeme naďalej uvažovať systém *zdanlivo nesúvisiacich regresíí* (3.7), kde $\mathbf{\Omega} = \mathbf{\Sigma} \otimes \mathbf{I} = \mathbf{E}\mathbf{e}_*\mathbf{e}'$ je neznáma a musíme ju odhadnúť. Pôvodne Zellner v roku 1962 navrhol nahradiť rozptyly σ_{ii} a kovariancie σ_{ij} konzistentnými odhadmi

$$s_{ij} = \frac{1}{T} \hat{\mathbf{e}}_i' \hat{\mathbf{e}}_j, \quad i, j = 2, \dots, M, \quad (3.12)$$

kde $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{y}_i$ reprezentuje *OLS* reziduá z i -tej rovnosti. Nech

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{M1} & s_{M2} & \dots & s_{MM} \end{bmatrix}$$

a $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Sigma}} \otimes \mathbf{I}$. *FGLS*-odhad systému *zdanlivo nesúvisiacich regresíí* je potom

$$\tilde{\boldsymbol{\beta}}_* = (\mathbf{X}_*'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}_*)^{-1}\mathbf{X}_*'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}_*.$$

Veta 3.7 Zellner. (*Dvojkrokový Zellnerov*) „*Feasible*“ odhad *zobecnenou metódou najmenších štvorcov* $\tilde{\boldsymbol{\beta}}_* = (\mathbf{X}_*'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}_*)^{-1}\mathbf{X}_*'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}_*$ má rovnaké asymptotické rozdelenie ako odhad *zobecnenou metódou najmenších štvorcov* $\tilde{\boldsymbol{\beta}}_* = (\mathbf{X}_*'\mathbf{\Omega}^{-1}\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{\Omega}^{-1}\mathbf{y}_*$. $\tilde{\boldsymbol{\beta}}_*$ je konzistentná a normálne rozdelená a navyše ak je \mathbf{e}_* normálne rozdelené, $\tilde{\boldsymbol{\beta}}_*$ je asymptoticky efektívny.

Poznamenajme, že asymptotické testy pre *zdanlivo nesúvisiace regresie* nevyžadujú normalitu chybovej zložky \mathbf{e}_* , avšak požadujú asymptotickú efektívnosť. Konzistentný odhad kovariančnej matice $\tilde{\boldsymbol{\beta}}_*$ je daný $(\mathbf{X}_*'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}_*)^{-1}$.

Veta 3.8 θ –rovnosti pre *zdanlivo nesúvisiace regresie* za predpokladu normality sú:

$$\mathbf{\Sigma} = \frac{1}{T} \mathbf{E}'\mathbf{E}, \quad (3.13)$$

kde $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_M]$. Informačná matica pre $\boldsymbol{\beta}_*$ a $\mathbf{\Sigma}$ je

$$\mathbf{I} = \begin{bmatrix} \mathbf{X}_*'\mathbf{\Omega}^{-1}\mathbf{X}_* & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\boldsymbol{\psi} \end{bmatrix},$$

kde $\boldsymbol{\psi}$ je matica $T \times T$ (definovaná Magnusom).

Pri konvergencii vedie ďalšia iterácia Zellnerovho dvojkrokového odhadu k Zellnerovmu iteračnému odhadu, ktorý je maximálne vierohodným odhadom systému *zdanlivo nesúvisiacich regresíí*.

Z hľadiska asymptotickej efektívnosti, ďalšou iteráciou už nič nezískame. Avšak stále zostáva nezodpovedaná otázka potenciálneho zisku z ďalšej iterácie na malých (konečných) výberoch.

3.1.5 Zdanlivo nesúvisiace regresie: výsledky na malom výbere

Vo všeobecnom prípade *FGLS* máme k dispozícii iba neustrannosť (ak existuje stredná hodnota) keď $\hat{\Omega}$ je funkcia \mathbf{e} . Zrejme nedostatočne špecifická štruktúra pre Ω je principiálnym dôvodom pre nedeterminovanie malého výberu. Dokonca ani predpoklad normálne rozdelených chýb konečne malého výberu nie je použiteľný, keď určitá charakteristická povaha Ω zostane nešpecifikovaná. Je niekoľko vzácných prípadov, kde v spojení s nejakým parametrickým rozdelením chýb, najčastejšie s normálnym rozdelením, vlastnosti *FGLS*-odhadu na malom výbere sú dostupné pre nejakú konkrétnu formu Ω . Jedným z týchto prípadov sú dve rovnice zdanlivo nesúvisiacich regresí s normálnymi chybami a s ortogonálnymi regresormi, $\mathbf{X}_1' \mathbf{X}_2 = \mathbf{0}$. Predpokladajme na chvíľu, že všetky regresory $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_M$ sú numericky odlišné. Uvažujme systém

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} (\mathbf{X}_1, \mathbf{X}_{D1}) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_2, \mathbf{X}_{D2}) & & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & (\mathbf{X}_M, \mathbf{X}_{DM}) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_{10} \\ \vdots \\ \beta_2 \\ \beta_{20} \\ \vdots \\ \vdots \\ \beta_M \\ \beta_{M0} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_M \end{bmatrix}, \quad (3.14)$$

kde β_i sú koeficienty pri \mathbf{X}_i a β_{i0} sú vektorové koeficienty rádu $\sum_{j \neq i} K_j \times 1$ zodpovedajúce $\mathbf{X}_{Di} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \dots, \mathbf{X}_M)$, maticiam rôznych regresorov zo všetkých rovníc s výnimkou i -tej. Odpovedajúci systém reziduí môžeme zapísať do tvaru

$$\hat{\mathbf{u}}_i = \hat{\mathbf{y}}_i - [\mathbf{X}_i \quad \mathbf{X}_{Di}] \begin{bmatrix} \hat{\beta}_i \\ \hat{\beta}_{i0} \end{bmatrix}, \quad (3.15)$$

kde $\hat{\beta}_i$ a $\hat{\beta}_{i0}$ označujú *OLS*-odhady β_i a β_{i0} .

Tieto reziduá môžeme požiť k výpočtu

$$s_{uij} = \frac{\hat{\mathbf{u}}_i' \hat{\mathbf{u}}_j}{T}, \quad i, j = 1, 2, \dots, M$$

a teda $MT \times MT$ matica

$$\hat{\Omega}_u = \hat{\Sigma}_u \otimes \mathbf{I} = \begin{bmatrix} s_{u11}\mathbf{I}_T & s_{u12}\mathbf{I}_T & \dots & s_{u1M}\mathbf{I}_T \\ s_{u21}\mathbf{I}_T & s_{u22}\mathbf{I}_T & & \\ \vdots & & \ddots & \vdots \\ s_{uM1}\mathbf{I}_T & s_{uM2}\mathbf{I}_T & \dots & s_{uMM}\mathbf{I}_T \end{bmatrix}. \quad (3.16)$$

Matica $\hat{\Sigma}_u$ sa nazýva *neobmedzený odhad* Σ a je konzistentným odhadom Σ . Odhad $\hat{\Sigma}$ môže byť odvodený z (3.13) so zavedením obmedzenia $\beta_{i0} = \mathbf{0}$ pre $i = 1, 2, \dots, M$.

Z tohto dôvodu sa $\hat{\Sigma}$ často nazýva *obmedzený odhad* Σ . Systém rovníc (3.14) sa dá ľahko modifikovať tak, aby niektoré (ale nie všetky) regresory boli rôzne.

Definícia 3.5 $\hat{\Sigma}_u$ je *neobmedzený odhad* Σ a $\hat{\Sigma}$ je *obmedzený odhad* Σ . Potom $\hat{\Omega}_u = \hat{\Sigma}_u \otimes \mathbf{I}$ a $\hat{\Omega} = \hat{\Sigma} \otimes \mathbf{I}$ sú *neobmedzený* a *obmedzený odhad* Ω , v tomto poradí.

Asymptoticky zrejme nezáleží na tom, ktorý z odhadov Σ použijeme pri konštrukcii *FGLS*-odhadu, avšak v prípade malého výberu to tak nutne nemusí byť.

Zellner vyšetroval na malom výbere vlastnosti neobmedzeného dvojkrovového odhadu

$$\tilde{\beta}_{*u} = (\mathbf{X}'_*(\hat{\Sigma}_u \otimes \mathbf{I})^{-1}\mathbf{X}_*)^{-1}\mathbf{X}'_*(\hat{\Sigma}_u \otimes \mathbf{I})^{-1}\mathbf{y}_* \quad (3.17)$$

v kontexte dvoch rovníc zdanlivo nesúvisiacich regresí s ortogonálnymi regresormi ($\mathbf{X}'_1\mathbf{X}_2 = 0$) a normálne rozdelenými chybami. Zellner ukázal, že $\tilde{\beta}_{*u}$ je nestranný. Ďalej vyšetroval relatívnu účinnosť *OLS*-odhadu β_1 , a to:

$$\hat{\beta}_1 = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{y}_1$$

k neobmedzenému odhadu $\tilde{\beta}_{1u}$. Zellner ukázal, že

$$\text{var}(\tilde{\beta}_{*u}) = \sigma_{11}(1 - \rho^2)[1 + 1/(T^* - 2)](\mathbf{X}'_1\mathbf{X}_1)^{-1},$$

kde $T^* = T - K_1 - K_2 (> 2)$ a $\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{1/2}$. Potom medzi $\text{var}(\tilde{\beta}_{1u})$ a $\text{var}(\hat{\beta}_1)$ je rovnaká (ne)rovnosť ako medzi $(1 - \rho^2)[1 + 1/(T^* - 2)]$ a 1, v tomto poradí. *FGLS*-odhad $\tilde{\beta}_{1u}$ je teda účinnejší ako *OLS*-odhad keď

$$\left(\frac{1}{T^* - 1}\right)^{1/2} < |\rho| < 1. \quad (3.18)$$

Z danej asymptotickej eficientnosti *FGLS*-odhadu $\tilde{\beta}_{1u}$, automaticky vyplýva, že $\tilde{\beta}_{1u}$ je účinnejší ako *OLS*-odhad $\hat{\beta}_1$ cez celý parametrický priestor, pokiaľ sa rozsah výberu blíži k nekonečnu. Ale pre malé výbery odhadnúť parametre σ_{11} , σ_{22} a σ_{12} je „drahé“, pokiaľ ρ je blízko pri nule a účinnejší odhad môže byť dosiahnutý ak apriori predpokladáme, že $\rho = 0$.

3.2 Heteroskedasticita

Za istých okolností býva porušená homoskedasticita chybovej zložky. Napríklad v priezovej analýze v ekonómii, sú obvyklým predmetom skúmania firmy, domácnosti alebo jednotlivci a to, do akej miery regresná rovnica vyjadruje ich chovanie môže závisieť na ich špecifických vlastnostiach.

Uvažujme príklad, v ktorom budeme vysvetľovať zisk firiem v danom roku v závislosti na ich investíciách. Predpokladajme, že pre daný rok a dané odvetvie je centrom záujmu zisk jednotlivých firiem v závislosti na marketingových nákladoch a nákladoch na výskum a vývoj. Čiže predpokladaná regresná rovnica v lineárnom prípade je

$$\Pi_i = a_1 + a_2 A_i + a_3 R_i + e_i, \quad i = 1, 2, \dots, n, \quad (3.19)$$

kde Π_i je zisk i -tej firmy, A_i sú marketingové náklady i -tej firmy, R_i sú náklady na výskum a vývoj v i -tej firme a n je počet skúmaných firiem.

Hoci je zrejmé logické, že stredné hodnoty chybových členov v (3.19) sú nulové, rozptyly chybových členov môžu byť u veľkých firiem väčšie ako u malých firiem, pretože investície malých firiem do výskumu a vývoja sú pravdepodobne podstatne menšie, zatiaľ čo veľké firmy majú väčší *leverage* (pomer cudzích zdrojov k vlastnému kapitálu) a teda aj väčšie investície. Takže u väčších firiem môžeme predpokladať väčšie riziko u návratnosti investícií, čiže väčší rozptyl okolo očakávaného zisku.

Ďalší zaujímavý príklad, kde sa rozptyl v závislosti na vysvetľovanej premennej mení uviedol Jacob Mincer v [10] v kapitole *Schooling, Experience, and Earnings*. Skúmal prírastky zárobkovej schopnosti jedincov v závislosti na počte rokov strávených v škole S a skúsenostiach (XPR) jedincov. Minceroва rovnica pre popis tejto závislosti je

$$\ln Y_i = \alpha_0 + \alpha_1 (XPR)_i - \alpha_2 (XPR)_i^2 + \alpha_3 S_i + \Phi_i' \alpha_4 + e_i,$$

kde Y_i je hodinový zárobok i -teho jedinca, $(XPR)_i$ sú skúsenosti i -teho jedinca mínus vzdelávanie v rokoch, S_i je vzdelávanie (v rokoch) i -teho jedinca a Φ_i je nejaký spoločný činiteľ ďalších premenných ako sú povolanie, rasa, odborár-neodborár, veľkosť mesta, priemyselné odvetvie, región. Pre nejakú konkrétnu úroveň vzdelania a ďalšie premenné sa rozptyly chýb v závislosti na skúsenostiach javia ako rýdzo konvexná funkcia v tvare U, t.j. chybový člen, a teda aj zárobky majú veľký rozptyl pri krátkych, ale aj dlhodobých skúsenostiach, naopak smerom do stredu sa rozptyly zmenšujú.

Samozrejme, zase existuje niekoľko logických dôvodov, prečo sú chyby na koncoch priezoru skúseností zväčšujú. Na vstupnej úrovni môže byť výrazný rozptyl spôsobený absolventskými zmluvami s nízkym nástupným platom s možnosťou výrazného platového rastu, a tiež absolventskými zmluvami s vysokým nástupným platom, ale len s malým následným rastom. Na druhej strane, pred ukončením kariéry, sa preddôchodkové plány jednotlivcov značne líšia. Niektorí jedinci odmietajú povýšenie, pretože sa nechcú sťahovať do iného mesta alebo nechcú stresujúcu prácu, prípadne nechcú pracovať na manažérskej pozícii. Pričom sloboda ich voľby závisí hlavne na ich rokmi nadobudnutom majetku. Taktiež predčasné dôchodky a produktivita v závislosti na veku a zdraví môže spôsobovať variabilitu príjmov.

V praxi nastáva mnoho prípadov, keď môžeme očakávať, že rozptyl chybovej zložky sa bude meniť (v závislosti na nejakom regresore, resp. na čase u časových radov). Na reálnych dátach však najčastejšie príčinu heteroskedasticity nepoznáme.

Z Aitkenovej vety je vidieť, že *OLS*-odhad je nestranný, ale nie je eficientný. Miesto $\mathbf{e}'\mathbf{e}$ môžeme uvažovať vhodnú stratovú funkciu $\mathbf{e}'\mathbf{\Omega}^{-1}\mathbf{e}$. V tejto časti budeme skúmať rôzne formy heteroskedasticity a diskutovať implementáciu týchto informácií do *GLS*-odhadov tak aby sme okrem nestrannosti odhadov dostali aj narastajúcu účinnosť.

Mohli by sme predpokladať, že heteroskedasticita je špecifikovaná všeobecnou formou

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_T^2 \end{bmatrix},$$

avšak v *GLS*-odhade použiť takto obecnú formu $\mathbf{\Omega}$ nie je možné, nakoľko počet neznámych parametrov je $(T + K)$, konkrétne $\beta_1, \beta_2, \dots, \beta_K$, a $\sigma_1^2, \sigma_2^2, \dots, \sigma_T^2$, ale k dispozícii máme iba T pozorovaní. Jedná sa o neriešiteľnú úlohu, takže musíme vyžadovať nejaké ďalšie obmedzenia na $\mathbf{\Omega}$.

3.2.1 Odhady v rôznych formách heteroskedasticity

I Skupinová heteroskedasticita

V mnohých prípadoch je možné rozpoznať skupiny pozorovaní s rovnakým rozptylom chýb v rámci jednotlivých skupín, ale s rôznymi rozptylmi medzi skupinami. Napr. pri skúmaní výdavkov na bývanie môžeme pozorovania rozdeliť do skupín podľa príjmu. Uvažujme napríklad T pozorovaní vhodných pre regresný model a predpokladajme, že existuje G skupín pozorovaní závislej premennej s T_1, T_2, \dots, T_G pozorovaniami v každej skupine ($T_1 + T_2 + \dots + T_G = T$). Ďalej predpokladajme, že pre všetky g je $T_g > K$, kde K = počet nezávislých premenných v lineárnom modeli. V maticovej forme:

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta} + \mathbf{e}_1, \\ \mathbf{y}_2 &= \mathbf{X}_2\boldsymbol{\beta} + \mathbf{e}_2, \\ &\vdots \\ \mathbf{y}_G &= \mathbf{X}_G\boldsymbol{\beta} + \mathbf{e}_G, \end{aligned} \tag{3.20}$$

kde \mathbf{y}_g , a \mathbf{X}_g sú pozorované závislé a nezávislé premenné v g -tej skupine a \mathbf{e}_g sú náhodné chyby v g -tej skupine. Koefficienty $\boldsymbol{\beta}' = (\beta_1, \beta_2, \dots, \beta_K)$ sú rovnaké v každej skupine. Redukovaný zápis pre *skupinovo heteroskedastický model* je

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{3.21}$$

kde $\mathbf{y}' = (\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_G)$, $\mathbf{e}' = (\mathbf{e}'_1, \mathbf{e}'_2, \dots, \mathbf{e}'_G)$ a $\mathbf{X}' = (\mathbf{X}'_1, \mathbf{X}'_2, \dots, \mathbf{X}'_G)$. Pre chyby v skupinovo heteroskedastickom modeli predpokladáme

$$\mathbf{E}\mathbf{e} = \mathbf{0} \quad (3.22)$$

a

$$\mathbf{E}\mathbf{e}\mathbf{e}' = \mathbf{E} \begin{bmatrix} \mathbf{e}_1\mathbf{e}'_1 & \mathbf{e}_1\mathbf{e}'_2 & \dots & \mathbf{e}_1\mathbf{e}'_G \\ \mathbf{e}_2\mathbf{e}'_1 & \mathbf{e}_2\mathbf{e}'_2 & \dots & \mathbf{e}_2\mathbf{e}'_G \\ & & \ddots & \\ \mathbf{e}_G\mathbf{e}'_1 & \mathbf{e}_G\mathbf{e}'_2 & \dots & \mathbf{e}_G\mathbf{e}'_G \end{bmatrix} = \begin{bmatrix} \sigma_1^2 \mathbf{I}_{T_1} & & & \mathbf{0} \\ & \sigma_2^2 \mathbf{I}_{T_2} & & \\ & & \ddots & \\ \mathbf{0} & & & \sigma_G^2 \mathbf{I}_{T_G} \end{bmatrix}, \quad (3.23)$$

kde $\mathbf{E}\mathbf{e}_g\mathbf{e}'_g = \sigma_g^2 \mathbf{I}_{T_g}$, $g = 1, 2, \dots, G$ a σ_g^2 je homoskedastický rozptyl v g -tej skupine.

V prípade, že rozptyly σ_g^2 sú známe, jednoducho použijeme zobecnenú metódu najmenších štvorcov. *GLS*-transformácia je v tvare

$$\mathbf{P}' = \begin{bmatrix} \frac{1}{\sigma_1} \mathbf{I}_{T_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{\sigma_G} \mathbf{I}_{T_G} \end{bmatrix}. \quad (3.24)$$

Použitím tejto transformačnej matice dostaneme transformované rovnice

$$\mathbf{P}'\mathbf{y} = \mathbf{P}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}'\mathbf{e}$$

alebo

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^*, \quad (3.25)$$

kde $\mathbf{y}^* = \mathbf{P}'\mathbf{y}$, $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$ a $\mathbf{e}^* = \mathbf{P}'\mathbf{e}$. Chyby z rovnice (3.25) sú homoskedastické v tom, že $\mathbf{E}\mathbf{e}^*\mathbf{e}^{*\prime} = \mathbf{I}_T$. Aplikáciou *OLS* na transformovanú rovnicu (3.25) je ekvivalentná s *GLS* aplikovanou na (3.21). Výsledný odhad parametru $\boldsymbol{\beta}$ je konzistentný najlepší nestranný lineárny odhad. Ak majú chyby navyše normálne rozdelenie, je dokonca nevychýleným odhadom s minimálnym rozptylom.

Za mnohých okolností je rozptyl σ_g^2 neznámy a musíme ho odhadnúť. Nech $\hat{\boldsymbol{\beta}}_g = (\mathbf{X}'_g\mathbf{X}_g)^{-1}\mathbf{X}'_g\mathbf{y}_g$ označuje odhad $\boldsymbol{\beta}$ metódou najmenších štvorcov aplikovanou na g -tú skupinu pozorovaní. Potom konzistentný nestranný odhad rozptylu chýb σ_g^2 v g -tej skupine je

$$\hat{\sigma}_g^2 = \frac{(\mathbf{y}_g - \mathbf{X}_g\hat{\boldsymbol{\beta}}_g)'(\mathbf{y}_g - \mathbf{X}_g\hat{\boldsymbol{\beta}}_g)}{T_g - K}. \quad (3.26)$$

Nech

$$\hat{\mathbf{P}}' = \begin{bmatrix} \frac{1}{\hat{\sigma}_1} \mathbf{I}_{T_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{\hat{\sigma}_G} \mathbf{I}_{T_G} \end{bmatrix},$$

tak potom odhad

$$\check{\boldsymbol{\beta}} = [(\hat{\mathbf{P}}'\mathbf{X})'(\hat{\mathbf{P}}'\mathbf{X})]^{-1}(\hat{\mathbf{P}}'\mathbf{X})\hat{\mathbf{P}}'\mathbf{y} \quad (3.27)$$

je *FGLS*-odhad v skupinovo heteroskedastickom modeli. Tento odhad je konzistentný a asymptoticky má normálne rozdelenie. Testovanie hypotéz ohľadom $\boldsymbol{\beta}$ je

založené na asymptotickej normalite a tom, že asymptotická kovariančná matica je v tomto prípade odhadovaná ako

$$[(\hat{\mathbf{P}}'\mathbf{X})'(\hat{\mathbf{P}}'\mathbf{X})]^{-1}. \quad (3.28)$$

V prípade, že chyby majú normálne rozdelenie, odhad je tiež asymptoticky efektívny.

II Heteroskedasticita ako funkcia exogénnych veličín

Heteroskedasticita môže byť často vyjadrená ako funkcia jednej alebo viacerých exogénnych veličín.

- (a) Pripomeňme si Mincerovu rovnicu: $\sigma_i^2 = a + b(XPR)_i + c(XPR)_i^2$. Tento rozptyl je funkciou exogénnych premenných.
- (b) Prais a Houthakker zistili, že rozptyly ich spotreby čaju sa proporcionálne rovná štvorcom stredných hodnôt regresnej funkcie:

$$\sigma_i^2 = \sigma^2(\mathbf{E}y_i)^2 = \sigma^2(\beta_1 x_{i1} + \dots + \beta_K x_{iK})^2,$$

kde σ^2 je neznáma konštanta. Ekvivalentne môžeme túto rovnicu prepísať bez druhých mocnín: $\sigma_i = \sigma(\beta_1 x_{i1} + \dots + \beta_K x_{iK}) = \beta_1^* x_{i1} + \dots + \beta_K^* x_{iK}$. V tomto prípade je smerodajná odchýlka lineárnou funkciou exogénnych veličín.

- (c) Pripomeňme si ešte prvý príklad zo sekcie o heteroskedasticite. Profit firmy v danom roku sme popisovali pomocou investícií firmy do marketingu, výskumu a vývoja. Rozptyl bol priamo závislý na nákladoch na výskum a vývoj. Tento vzťah môže byť reprezentovaný ako $\sigma_i^2 = \sigma^2(R_i)^\lambda$, kde σ^2 a λ sú neznáme reálne konštanty. Takže rozptyl chýb je multiplikatívny v parametri λ .

Tieto príklady rozhodne nepopisujú všetky typy heteroskedasticity, ale ukazujú veľmi časté typy, ktorými sú

$$\sigma_i^2 = \mathbf{z}_i' \boldsymbol{\alpha}, \quad (3.29)$$

$$\sigma_i^2 = (\mathbf{z}_i' \boldsymbol{\alpha})^2 \quad (3.30)$$

$$\sigma_i^2 = \exp(\mathbf{z}_i' \boldsymbol{\alpha}), \quad (3.31)$$

kde $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ je vektor pozorovaní na množine exogénnych premenných, ktoré obvykle, ale nie nutne, súvisia s regresormi a $\boldsymbol{\alpha}' = (\alpha_1, \alpha_2, \dots, \alpha_p)$ je vektor parametrov typu $p \times 1$, pričom vždy predpokladáme, že $z_{i1} \equiv 1$ pre všetky i . Pozrime sa ďalej aj na odhady lineárnych modelov so spomínanými typmi heteroskedasticity.

(A) $\sigma_i^2 = \mathbf{z}_i' \boldsymbol{\alpha}$

Lineárny model môže byť v tomto prípade vyjadrený v tvare

$$\mathbf{y}_i = \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{e}_i, \quad i = 1, 2, \dots, T, \quad (3.32)$$

pričom $\mathbf{E}(\mathbf{e}_i) = \mathbf{0}$, $\mathbf{E}(\mathbf{e}_i^2) = \sigma_i^2 = \mathbf{z}_i' \boldsymbol{\alpha}$, $\mathbf{E}(\mathbf{e}_i \mathbf{e}_j) = \mathbf{0}$, $i \neq j$. Navyše predpokladajme, že chyby \mathbf{e}_i majú normálne rozdelenie. Nech \mathbf{D}_1 je diagonálna matica s i -tým diagonálnym prvkom $\mathbf{z}_i' \boldsymbol{\alpha}$. GLS-odhad parametru $\boldsymbol{\beta}$ je

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{D}_1^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{D}_1^{-1} \mathbf{y} = \left(\sum_{i=1}^T (\mathbf{z}_i' \boldsymbol{\alpha})^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^T (\mathbf{z}_i' \boldsymbol{\alpha})^{-1} \mathbf{x}_i y_i$$

so strednou hodnotou β a kovariančnou maticou

$$(\mathbf{X}'\mathbf{D}_1^{-1}\mathbf{X})^{-1} = \left(\sum_{i=1}^T (\mathbf{z}'_i \boldsymbol{\alpha})^{-1} \mathbf{x}_i \mathbf{x}'_i \right)^{-1}.$$

Vektorový parameter $\boldsymbol{\alpha}$ je väčšinou neznámy a musíme ho odhadnúť, aby sme mohli použiť *GLS*.

Intuitívne by sme mohli nahradiť σ_i^2 v rovnici $\sigma_i^2 = \mathbf{z}'_i \boldsymbol{\alpha}$ konzistentným (*OLS*) odhadom $\hat{\mathbf{e}}_i^2$, t.j. $\hat{\mathbf{e}}_i^2 = (\mathbf{y}_i - \mathbf{x}_i \hat{\boldsymbol{\beta}})^2$, kde $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ je *OLS*-odhad parametru $\boldsymbol{\beta}$. Výsledná rovnica je potom

$$\hat{e}_i^2 = \mathbf{z}'_i \boldsymbol{\alpha} + \nu_i, \quad i = 1, 2, \dots, T, \quad (3.33)$$

kde

$$\nu_i = \hat{e}_i^2 - \sigma_i^2.$$

Avšak

$$\mathbf{E}\nu_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_1\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{d} \neq 0,$$

kde $\mathbf{d} = (0, \dots, 0, \sigma_i^2, 0, \dots, 0)$ je vektor typu $1 \times T$ s jediným nenulovým prvkom σ_i^2 na i -tom mieste. Keďže ν_i má nenulovú strednú hodnotu, ktorá sa mení s i , *OLS*-odhad $\boldsymbol{\alpha}$ je na malých výberoch vychýlený, ale ak je rozsah výberu dostatočne veľký obdržíme konzistentný odhad $\boldsymbol{\alpha}$ (stačí uvažovať $\text{plim}(\hat{e}_i - e_i)$).

Ak je rozsah výberu dostatočne veľký, *OLS* aplikovaná na \hat{e}_i^2 môže prinajmenšom poskytnúť konzistentný odhad $\boldsymbol{\alpha}$.

Goldfeld a Quandt v roku 1972 navrhli *OLS*-odhad $\boldsymbol{\alpha}$ spočítať z (3.33):

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^T \mathbf{z}_i \mathbf{z}'_i \right)^{-1} \sum_{i=1}^T \mathbf{z}_i \hat{e}_i^2 = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}}^2, \quad (3.34)$$

kde $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T)$ a $\hat{\mathbf{e}}^2 = (\hat{e}_1^2, \hat{e}_2^2, \dots, \hat{e}_T^2)$. Označme $\boldsymbol{\Sigma}_1$ diagonálnu maticu s i -tým diagonálnym prvkom $2\sigma_i^4$. Potom $\hat{\boldsymbol{\alpha}}$ je asymptoticky normálne so strednou hodnotou $\boldsymbol{\alpha}$ a s kovariančnou maticou $\mathbf{V}_1 = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_1\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$.

FGLS-odhad s použitím $\hat{\boldsymbol{\alpha}}$ je potom

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{D}}_1^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{D}}_1^{-1}\mathbf{y} = \left(\sum_{i=1}^T (\mathbf{z}'_i \hat{\boldsymbol{\alpha}})^{-1} \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \sum_{i=1}^T (\mathbf{z}'_i \hat{\boldsymbol{\alpha}})^{-1} \mathbf{x}_i y_i, \quad (3.35)$$

kde $\hat{\mathbf{D}}_1$ je diagonálna matica s i -tým diagonálnym prvkom $\mathbf{z}'_i \hat{\boldsymbol{\alpha}}$. Tento odhad je konzistentný, asymptoticky normálny a pokiaľ chyby e_i majú normálne rozdelenie, je aj asymptoticky efektívny. Konzistentný odhad jeho asymptotickej kovariančnej matice je $(\mathbf{X}'\hat{\mathbf{D}}_1^{-1}\mathbf{X})^{-1}$.

Amemiya v roku 1977 navrhol nasledujúcu trojkrokovú procedúru pre odhadnutie $\boldsymbol{\alpha}$:

1. spočítame *OLS*-odhad $\hat{\alpha}$,
2. skonštruujeme diagonálnu maticu $\hat{\Sigma}_1$, s i -tým diagonálnym prvkom $2\hat{\sigma}_i^4 = 2(\mathbf{z}_i' \hat{\alpha})^2$,
3. skonštruujeme *FGLS*-odhad α :

$$\hat{\alpha} = (\mathbf{Z}' \hat{\Sigma}_1^{-1} \mathbf{Z})^{-1} \mathbf{Z}' \hat{\Sigma}_1^{-1} \hat{\mathbf{e}}^2 = \left(\sum_{i=1}^T (\mathbf{z}_i' \hat{\alpha})^{-1} \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i=1}^T (\mathbf{z}_i' \hat{\alpha})^{-1} \mathbf{z}_i \hat{e}_i^2 \quad (3.36)$$

a ukázal, že tento odhad je asymptoticky normálny s nulovou strednou hodnotou a kovariančnou maticou

$$\mathbf{V}_2 = (\mathbf{Z}' \hat{\Sigma}_1^{-1} \mathbf{Z})^{-1}.$$

V prípade, že chyby majú normálne rozdelenie, tento odhad je asymptoticky efektívnym odhadom α . $\hat{\alpha}$ však nie je asymptoticky efektívny, pretože matica $\mathbf{V}_1 - \mathbf{V}_2$ je pozitívne semidefinitná, čo nastáva, pretože $\hat{\alpha}$ neberie v úvahu heteroskedasticitu.

Hoci nemáme k dispozícii žiadne analytické vlastnosti na malých výberoch, zdá sa byť rozumné, aby *FGLS*-odhadom pre heteroskedasticitu typu $\sigma_i^2 = \mathbf{z}_i' \alpha$ bolo $\tilde{\beta}$:

$$\tilde{\beta} = (\mathbf{X}' \hat{\mathbf{D}}_1^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{D}}_1^{-1} \mathbf{y}, \quad (3.37)$$

kde $\hat{\mathbf{D}}$ reprezentuje diagonálnu maticu s i -tým diagonálnym prvkom $\mathbf{z}_i' \hat{\alpha}$. Podobne ako *FGLS*-odhad, aj tento odhad je konzistentný, má asymptoticky normálne rozdelenie a ak majú chyby tiež normálne rozdelenie, je aj asymptoticky efektívny. A konečne asymptotická kovariančná matica pre tento odhad je konzistentne odhadnutá ako $(\mathbf{X}' \hat{\mathbf{D}}_1^{-1} \mathbf{X})^{-1}$.

(B) $\sigma_i = (\mathbf{z}_i' \alpha)$

Budeme sa zaoberať lineárnym modelom v tvare

$$y_i = \mathbf{x}_i' \beta + e_i, \quad i = 1, 2, \dots, T, \quad (3.38)$$

pričom $E(e_i) = 0$, $E(e_i^2) = \sigma_i^2 = (\mathbf{z}_i' \alpha)^2$, $E(e_i e_j) = 0$, $i \neq j$ a e_i majú normálne rozdelenie. Nech \mathbf{D}_2 je diagonálna matica s i -tým diagonálnym prvkom $(\mathbf{z}_i' \alpha)^2$. *GLS*-odhad parametru β je

$$\tilde{\beta} = (\mathbf{X}' \mathbf{D}_2^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{D}_2^{-1} \mathbf{y} = \left(\sum_{i=1}^T (\mathbf{z}_i' \alpha)^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^T (\mathbf{z}_i' \alpha)^{-2} \mathbf{x}_i y_i$$

so strednou hodnotou β a kovariančnou maticou

$$(\mathbf{X}' \mathbf{D}_2^{-1} \mathbf{X})^{-1} = \left(\sum_{i=1}^T (\mathbf{z}_i' \alpha)^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1}.$$

Diskusia ohľadom *FGLS*-odhadu bude obdobná ako v predchádzajúcej sekcii až na odhad pre α , ktorý bude iný. Za predpokladu, že $\sigma_1^{-1}e_1, \sigma_2^{-1}e_2, \dots, \sigma_T^{-1}e_T$ sú *i.i.d.* s nulovou strednou hodnotou a jednotkovým rozptylom. Stredná hodnota absolútnych hodnôt týchto náhodných veličín je:

$$E\Phi = E(\sigma_i^{-1}|e_i|) = 2 \int_0^\infty w f(w) dw \equiv c,$$

kde $f(w)$ je hustota pravdepodobnosti $w = e_i/\sigma_i$ a c je konštanta, ktorá je nezávislá na i a závisí na špecifickej hustote $f(w)$, o ktorej predpokladáme, že je symetrická okolo nuly a definovaná na doméne reálnych čísel (rovnako ako e_i). Takže ak napríklad $e_i \sim N(0, \sigma_i^2)$, potom $w = e_i/\sigma_i \sim N(0, 1)$ zo základných vlastností normálneho rozdelenia, viď napr. [3]. Takže strednú hodnotu ľahko vypočítame:

$$E\Phi = 2 \int_0^\infty w \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{w^2}{2}\right\} dw = \frac{2}{\sqrt{2\pi}} \int_0^\infty w \exp\left\{-\frac{w^2}{2}\right\} dw = \sqrt{\frac{2}{\pi}} \equiv c.$$

Teda $e|e_i| = c\sigma_i^2 = c\mathbf{z}_i'\alpha$ a lineárny odhad dostaneme logicky z

$$|\hat{e}_i| = c\mathbf{z}_i'\alpha + \nu_i,$$

kde $\hat{e}_i = y_i - \mathbf{x}_i'\hat{\beta}$ je *OLS*-odhad i -teho rezidua a $\nu_i = |\hat{e}_i| - c\sigma_i$. Gleisjer v roku 1969 navrhol odhadnúť $c\alpha$ pomocou *OLS*, čím dostaneme:

$$\hat{c}\alpha = \left(\sum_{i=1}^T \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i=1}^T \mathbf{z}_i |\hat{e}_i| = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'|\hat{\mathbf{e}}|, \quad (3.39)$$

kde $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T)'$ a $|\hat{\mathbf{e}}| = (|\hat{e}_1|, |\hat{e}_2|, \dots, |\hat{e}_T|)'$. To, že miesto odhadu α použijeme odhad $c\alpha$ nemá vplyv na výsledný *FGLS*-odhad β a jeho vlastnosti. Zodpovedajúci *FGLS*-odhad je potom:

$$\tilde{\beta} = \left(\sum_{i=1}^T (\mathbf{z}_i' \hat{c}\alpha)^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^T (\mathbf{z}_i' \hat{c}\alpha)^{-2} \mathbf{x}_i y_i. \quad (3.40)$$

Odhad $\hat{c}\alpha$ je však na malých výberoch vychýlený, nie je eficientný a ν_i sú heteroskedastické, autokorelované a majú rôzne a nenulové stredné hodnoty. Analogicky ako v predchádzajúcom prípade, chyby ν_i majú asymptoticky nulovú strednú hodnotu a sú nezávislé a heteroskedastické. Preto aj odhady $|\hat{e}_i|$ sú heteroskedastické a *OLS*-odhad $\hat{c}\alpha$ nie je asymptoticky eficientný odhad $c\alpha$. Lepšou variantou je *FGLS*-odhad $c\alpha$, ktorý získame nasledujúcim algoritmom:

1. Spočítame *OLS*-odhad $\hat{c}\alpha$, pre známe c , $\hat{\alpha} = \hat{c}\alpha/c$
2. Skonstruujeme konzistentný odhad kovariančnej matice odhadovaných chýb, Σ_2 , ktorú označíme $\hat{\Sigma}_2$. Jedná sa o diagonálnu maticu s i -tým diagonálnym prvkom $(1 - c^2)(\mathbf{z}_i' \hat{\alpha})^2$.

3. Skonstruujeme *FGLS*-odhad $c\alpha$:

$$\hat{c}\alpha = (\mathbf{Z}'\hat{\Sigma}_2^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\hat{\Sigma}_2^{-1}|\mathbf{e}|. \quad (3.41)$$

Zodpovedajúci *FGLS*-odhad β je potom

$$\tilde{\beta} = \left(\sum_{i=1}^T (\mathbf{z}_i' \hat{c}\alpha)^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^T (\mathbf{z}_i' \hat{c}\alpha)^{-2} \mathbf{x}_i y_i. \quad (3.42)$$

Pokiaľ majú chyby e_i normálne rozdelenie, oba *FGLS*-odhady sú asymptoticky efficientné.

(C) $\sigma_i^2 = \exp(\mathbf{z}_i' \alpha)$

V roku 1976 Harvey navrhol a skúmal všeobecnú formuláciu heteroskedasticity - multiplikatívnu heteroskedasticitu. Lineárny model špecifikoval

$$y_i = \mathbf{x}_i' \beta + e_i, \quad i = 1, 2, \dots, T, \quad (3.43)$$

pričom $E(e_i) = 0$, $E(e_i^2) = \sigma_i^2 = \exp(\mathbf{z}_i' \alpha)$, $E(e_i e_j) = 0$, $i \neq j$ a e_i majú normálne rozdelenie.

Nech \mathbf{D}_3 je diagonálna matica s i -tým diagonálnym prvkom $\exp(\mathbf{z}_i' \alpha)$. *GLS*-odhad parametru β je

$$\begin{aligned} \tilde{\beta} &= (\mathbf{X}' \mathbf{D}_3^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{D}_3^{-1} \mathbf{y} \\ &= \left(\sum_{i=1}^T [\exp(\mathbf{z}_i' \alpha)]^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^T [\exp(\mathbf{z}_i' \alpha)]^{-1} \mathbf{x}_i y_i \end{aligned}$$

so strednou hodnotou β a kovariančnou maticou

$$(\mathbf{X}' \mathbf{D}_3^{-1} \mathbf{X})^{-1} = \left(\sum_{i=1}^T [\exp(\mathbf{z}_i' \alpha)]^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1}.$$

Väčšinou musíme parametrický vektor α odhadnúť konzistentne, aby sme mohli použiť *FGLS*.

Aplikovaním logaritmickkej transformácie na $\sigma_i^2 = \exp(\mathbf{x}_i' \alpha)$ dostaneme

$$\ln \sigma_i^2 = \mathbf{z}_i' \alpha. \quad (3.44)$$

Obdobne ako v predchádzajúcom prípade nahradíme σ_i^2 odhadom \hat{e}_i^2 : $\ln \hat{e}_i^2 = \mathbf{z}_i' \alpha + \nu_i$, kde $\nu_i = \ln(\hat{e}_i^2 / \sigma_i^2)$ a \hat{e}_i je *OLS*-odhad reziduí. Aplikáciou metódy najmenších štvorcov na (3.44) dostaneme:

$$\hat{\alpha} = \left(\sum_{i=1}^T \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i=1}^T \mathbf{z}_i \ln \hat{e}_i^2. \quad (3.45)$$

FGLS-odhad parametru β má tvar

$$\tilde{\beta} = \left(\sum_{i=1}^T [\exp(\mathbf{z}_i' \hat{\alpha})] \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^T [\exp(\mathbf{z}_i' \hat{\alpha})] \mathbf{x}_i y_i \quad (3.46)$$

je skonštruovaný z $\hat{\alpha}$, je konzistentný, a má asymptoticky normálne rozdelenie a ak majú chyby tiež normálne rozdelenie, je aj asymptoticky eficientný. Harvey tiež ukázal, že asymptotická kovariančná matica odhadu $\hat{\alpha}$ je

$$\text{var}(\hat{\alpha}) = 4.9348 \left[\sum_{i=1}^T \mathbf{z}_i \mathbf{z}_i' \right]^{-1} \quad (3.47)$$

Heteroskedasticita ako známa, ale všeobecná funkcia

Predchádzajúcimi tromi prípadmi sme rozhodne nepopísali všetky typy heteroskedasticity, preto vyjadrenie heteroskedasticity ešte zovšeobecníme.

Nech $f(\mathbf{z})$ označuje nejakú prostú funkciu v premennej $\mathbf{z} = (z_1, z_2, \dots, z_p)$ na kladnej časti množiny reálnych čísel. Predpokladajme heteroskedasticitu v tvare

$$\sigma_i^2 = f(\mathbf{z}_i), \quad (3.48)$$

kde $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ označuje i -te pozorovanie z p exogénnych premenných z_1, z_2, \dots, z_p a $i = 1, 2, \dots, T$. Napríklad σ_i^2 môže mať tvar

$$\sigma_i^2 = f(\mathbf{z}_i) = \frac{a + bz_{i1}}{c + dz_{i2}},$$

pričom predmetom záujmu je opäť lineárny model $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, s centrovanými chybami, $\mathbf{E}\mathbf{e} = \mathbf{0}$, a s kovariančnou maticou chýb

$$\mathbf{\Omega} = \mathbf{E}\mathbf{e}\mathbf{e}' = \begin{bmatrix} f(\mathbf{z}_1) & & & 0 \\ & f(\mathbf{z}_2) & & \\ & & \ddots & \\ 0 & & & f(\mathbf{z}_T) \end{bmatrix}.$$

V prípade, že chyby sú parametrizované, môžeme použiť maximálne vierohodný odhad. Za predpokladu normálneho rozdelenia chýb vierohodnostná funkcia

$$L = \left(\frac{1}{\sqrt{2\pi}} \right)^T |\mathbf{\Omega}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \mathbf{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\beta) \right\}$$

maximalizuje parametre funkcie $f(\mathbf{z})$ vzhľadom k β .

3.2.2 Detekcia heteroskedasticity

V predchádzajúcej časti sme predpokladali, že heteroskedasticita je apriórne známa. V praxi však najčastejšie nepoznáme príčiny heteroskedasticity, môžeme mať síce podozrenie na jej formu, existenciou heteroskedasticity si ale nie sme istí. V takomto prípade môžeme použiť nejaký vhodný štatistický test. Ak sa heteroskedasticita potvrdí, musíme tomu prispôbiť aj nasledovné riešenie. V opačnom prípade môžeme použiť *OLS*. Špecifikácia chýb založená na testovaní hypotéz teda vedie k predbežnému výberu *OLS* alebo *FGLS*, preto niekedy hovoríme o predbežnom teste. V tejto časti popíšeme testy hypotéz, ktoré sú vhodné pre zmieňované typy heteroskedasticity, t.j. skupinovú heteroskedasticitu a heteroskedasticitu ako funkcia exogénnych premenných.

3.2.3 Test založený na pomere vierohodností pre skupinovú heteroskedasticitu

Nech je daný model skupinovej heteroskedasticity (3.21)–(3.23) a chceme testovať homoskedasticitu proti skupinovej heteroskedasticite:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_G^2 = \sigma^2 \\ H_1 &: \text{non } H_0 \end{aligned} \quad (3.49)$$

Predpokladajme, že e_i sú normálne rozdelené. Použijeme podiel vierohodností: $l = L(\hat{\omega})/L(\hat{\Omega})$ z vety 2.12, pričom môžeme použiť asymptotický test (3.49). Za platnosti nulovej hypotézy $-2 \ln l \sim \chi_q^2$. Ďalej Ω je $(K + G)$ –rozmerná všeobecná parametrická množina parametrov $\{\beta_1, \beta_2, \dots, \beta_K, \sigma_1^2, \dots, \sigma_G^2\}$ a ω je $[(K + G) - (G - 1)]$ –rozmerná podmnožina Ω , s prvkami $\{\beta_1, \dots, \beta_K, \sigma^2\}$. Takže potrebujeme zostrojiť testovú štatistiku l . Nakoľko parametre modelu nepoznáme, nahradíme ich maximálne vierohodnými odhadmi, t.j. v prípade reštrikcie na množinu ω použijeme miesto parametru β jeho odhad $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ a miesto σ^2 použijeme $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})/T$. Pokiaľ chceme vierohodnostnú funkciu zostrojiť na celom parametrickom priestore, použijeme odhad parametru β v tvare $\tilde{\beta} = (\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^{-1}\mathbf{y}$, kde

$$\mathbf{S} = \begin{bmatrix} \tilde{\sigma}_1^2 \mathbf{I}_{T_1} & & & 0 \\ & \tilde{\sigma}_2^2 \mathbf{I}_{T_2} & & \\ & & \ddots & \\ 0 & & & \tilde{\sigma}_G^2 \mathbf{I}_{T_G} \end{bmatrix}$$

a $\tilde{\sigma}_g^2 = (\mathbf{y}_g - \mathbf{X}_g\tilde{\beta})(\mathbf{y}_g - \mathbf{X}_g\tilde{\beta})'/T_g$, $g = 1, 2, \dots, G$ pričom odhady na celom parametrickom priestore Ω dostaneme iteratívne, ako iniciálny použijeme odhad $\hat{\beta}_g = (\mathbf{X}_g'\mathbf{X}_g)^{-1}\mathbf{X}_g'\mathbf{y}_g$ a s iterovaním skončíme, keď rozdiely medzi dvoma po sebe idúcimi odhadmi budú zanedbateľne malé. Konečne, vierohodnostné funkcie na celom a zúženom parametrickom priestore budú vyzerat nasledovne:

$$L(\hat{\omega}) = \frac{1}{(2\pi)^{T/2}\tilde{\sigma}^T} \exp \left\{ -\frac{1}{2\tilde{\sigma}^2}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) \right\} \quad (3.50)$$

a

$$L(\hat{\Omega}) = \frac{1}{(2\pi)^{T/2} \tilde{\sigma}_1^{T_1} \tilde{\sigma}_2^{T_2} \dots \tilde{\sigma}_G^{T_G}} \exp \left\{ -\frac{1}{2} \sum_{g=1}^G \frac{1}{\tilde{\sigma}_g^2} (\mathbf{y}_g - X_g \tilde{\boldsymbol{\beta}})' (\mathbf{y}_g - X_g \tilde{\boldsymbol{\beta}}) \right\}. \quad (3.51)$$

Takže výsledný pomer vierohodností je

$$l = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{\tilde{\sigma}_1^{T_1} \tilde{\sigma}_2^{T_2} \dots \tilde{\sigma}_G^{T_G}}{\tilde{\sigma}^T} \quad (3.52)$$

a testová štatistika

$$-2 \ln l = 2[T \ln \tilde{\sigma} - (T_1 \ln \tilde{\sigma}_1 + T_2 \ln \tilde{\sigma}_2 + \dots + T_G \ln \tilde{\sigma}_G)] \quad (3.53)$$

má asymptoticky χ_{G-1}^2 rozdelenie.

3.2.4 Test heteroskedasticity ako funkcie exogénnych premenných

Test rozptylu ako lineárnej funkcie exogénnych premenných: $\sigma_i^2 = \mathbf{z}_i' \boldsymbol{\alpha}$

Opäť sa budeme chvíľu zaoberať heteroskedasticitou v tvare

$$\sigma_i^2 = \mathbf{z}_i' \boldsymbol{\alpha} \quad (3.54)$$

a budeme testovať nulovú hypotézu homoskedasticity, proti alternatíve, že rozptyl chýb má tvar (3.54), t.j.:

$$\begin{aligned} H_0 &: \alpha_2 = \alpha_3 = \dots = \alpha_p = \alpha \\ H_1 &: \text{non } H_0 \end{aligned} \quad (3.55)$$

Existujú dve rôzne možnosti riešenia:

- I. Pomocou bežnej metódy najmenších štvorcov odhadneme (tak ako sme popísali v (3.34)) parameter $\boldsymbol{\alpha}$, čím dostane *OLS*-odhad

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^p \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i=1}^p \mathbf{z}_i \hat{e}_i^2 = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{e}^2,$$

ktorý je asymptoticky normálny a jeho kovariančná matica má tvar

$$\mathbf{V}_1 = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \boldsymbol{\Sigma}_1 \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1},$$

kde $\boldsymbol{\Sigma}_1 = \text{diag}\{2\sigma_i^4\}$. Vidíme, že ak platí nulová hypotéza, kovariančná matica má tvar $\mathbf{V}_1 = 2\sigma_1^2 (\mathbf{Z}' \mathbf{Z})^{-1}$. Označme $\alpha_0 = (\hat{\alpha}_2, \hat{\alpha}_3, \dots, \hat{\alpha}_p)$ a \mathbf{V}_1^0 maticu $(p-1) \times (p-1)$, ktorá vznikne z matice \mathbf{V}_1 za H_0 , t.j. z matice $2\sigma_1^2 (\mathbf{Z}' \mathbf{Z})^{-1}$ vynechaním prvého riadku aj stĺpca. A konečne, označme TS_1 , kvadratickú formu

$$\hat{\boldsymbol{\alpha}}_0' (\mathbf{V}_1^0)^{-1} \hat{\boldsymbol{\alpha}}_0, \quad (3.56)$$

ktorú použijeme ako testovú štatistiku. Za platnosti nulovej hypotézy H_0 , má TS_1 asymptoticky χ_{p-1}^2 rozdelenie.

- II. Ako alternatíva k prvému riešeniu pomocou *OLS* sa ponúka analogické riešenie s tým rozdielom, že parameter α odhadneme pomocou *FGLS*. Dostaneme odhad podľa (3.36), povedzme $\hat{\alpha}$, ktorý má opäť asymptoticky normálne rozdelenie s kovariančnou maticou

$$\mathbf{V}_2 = (\mathbf{Z}'\Sigma_1^{-1}\mathbf{Z})^{-1}.$$

Alternatívnu testovú štatistiku označíme TS_2 a vypočítame ju obdobne ako testovú štatistiku v predchádzajúcom prípade:

$$TS_2 = \hat{\alpha}_0'(\mathbf{V}_2^0)^{-1}\hat{\alpha}_0, \quad (3.57)$$

kde $\hat{\alpha}_0 = (\hat{\alpha}_2, \hat{\alpha}_3, \dots, \hat{\alpha}_p)'$ a \mathbf{V}_2^0 označuje maticu, ktorá vznikla z \mathbf{V}_2^0 za H_0 vynechaním prvého riadku a aj prvého stĺpca. Za H_0 je asymptotické rozdelenie testovej štatistiky TS_2 zasa χ_{p-1}^2 , rovnako ako to bolo u TS_1 . Takže máme k dispozícii dva asymptoticky ekvivalentné testy. V prípade, že nemáme k dispozícii žiadnu ďalšiu znalosť, javí sa ako rozumné použiť radšej testovú štatistiku TS_2 .

Test smerodajnej odchýlky ako lineárnej funkcie exogénnych premenných:
 $\sigma_i = \mathbf{z}_i'\alpha$

Podobne ako v predchádzajúcej časti, test homoskedasticity, tentokrát proti alternatíve $\sigma_i = \mathbf{z}_i'\alpha$, znamená testovať nulovú hypotézu $\alpha_2 = \alpha_3 = \dots = \alpha_p = 0$. Opäť môžeme použiť *OLS*-odhad parametru α podľa (3.39), $\hat{\alpha}$, alebo *FGLS*-odhad podľa (3.41), $\hat{\alpha}$, pre známe c . Zodpovedajúce kovariančné matice pre $\hat{\alpha}$, a $\hat{\alpha}$, v tomto poradí, sú:

$$\frac{1}{c^2}\mathbf{V}_3 = \frac{1}{c^2}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\Sigma}_2\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \quad (3.58)$$

a

$$\frac{1}{c^2}\mathbf{V}_4 = \frac{1}{c^2}(\mathbf{Z}'\hat{\Sigma}_2^{-1}\mathbf{Z})^{-1}, \quad (3.59)$$

kde $\hat{\Sigma}_2 = \text{diag}\{(1 - c^2)(\mathbf{z}_i'\hat{\alpha})^2\}$ a $\hat{\Sigma}_2 = \text{diag}\{(1 - c^2)(\mathbf{z}_i'\hat{\alpha})^2\}$. Potom zodpovedajúce testové štatistiky sú

$$TS_3 = \hat{\alpha}_0'(\mathbf{V}_3^0)^{-1}\hat{\alpha}_0 \quad (3.60)$$

a

$$TS_4 = \hat{\alpha}_0'(\mathbf{V}_4^0)^{-1}\hat{\alpha}_0, \quad (3.61)$$

kde $\hat{\alpha}_0 = (\hat{\alpha}_2, \hat{\alpha}_3, \dots, \hat{\alpha}_p)$, $\hat{\alpha}_0 = (\hat{\alpha}_2, \hat{\alpha}_3, \dots, \hat{\alpha}_p)$, a \mathbf{V}_3^0 a \mathbf{V}_4^0 sú $(p - 1) \times (p - 1)$ matice, ktoré vznikli vynechaním prvého riadku a aj prvého stĺpca v maticiach $[(1 - c^2)\hat{\alpha}_1^2/c^2](\mathbf{Z}'\mathbf{Z})^{-1}$ a $[(1 - c^2)\hat{\alpha}_1^2/c^2](\mathbf{Z}'\mathbf{Z})^{-1}$. Obe testové štatistiky, TS_3 aj TS_4 majú za platnosti nulovej hypotézy asymptoticky χ_{p-1}^2 rozdelenie.

Test multiplikatívnej heteroskedasticity: $\sigma_i = \exp(\mathbf{z}_i'\alpha)$

Multiplikatívny parameter α môžeme odhadnúť pomocou bežnej metódy najmenších štvorcov a pomocou (3.45) dostaneme *OLS* odhad $\hat{\alpha}$, konzistentný, ale neeficientný

odhad. Zasa rovnako ako v predchádzajúcich prípadoch, test bude založený na hypotéze, $\alpha_2 = \alpha_3 = \dots = \alpha_p = 0$. Nech \mathbf{V}_5 označuje kovariančnú maticu z 3.47 a \mathbf{V}_5^0 je matica, ktorá vznikne z \mathbf{V}_5 vynechaním prvého riadku a prvého stĺpca, takže je typu $(p-1) \times (p-1)$. Ďalej nech opäť $\hat{\boldsymbol{\alpha}}_0 = (\hat{\alpha}_2, \hat{\alpha}_3, \dots, \hat{\alpha}_p)'$. Potom ak je rozptyl σ_i konštantný, t.j. platí nulová hypotéza homoskedasticity, štatistika

$$TS_5 = \hat{\boldsymbol{\alpha}}_0'(\mathbf{V}_5^0)^{-1}\hat{\boldsymbol{\alpha}}_0 \quad (3.62)$$

má asymptotické rozdelenie χ_{p-1}^2 .

Test heteroskedasticity v známej ale všeobecnej forme: Metódy maximálnej vierohodnosti

Pre prípad, že ani jedna z troch foriem heteroskedasticity, ku ktorým sme doposiaľ skonštruovali testy sa nejaví ako vhodná, potrebujeme riešenie pre všeobecnejšiu formu heteroskedasticity. Nech platí, že $\sigma_i^2 = f(\mathbf{z}_i)$, čiže heteroskedasticita je popísaná nejakou funkciou $f(\mathbf{z}_i)$, ktorá je ale odlišná od už spomínaných foriem heteroskedasticity. V prípade, že poznáme distribúciu chybovej zložky e_i , môžeme použiť maximálnu vierohodnosť a homoskedasticitu proti heteroskedasticite testovať pomocou testov založených na vierohodnostnej funkcii.

Okrem testu založenom na pomere vierohodností môžeme alternatívne použiť *test založený na Lagrangeových multiplikátoroch* (známy tiež pod názvami *skórový test* alebo *Raov test*) alebo *Waldov test*, ktoré tiež vychádzajú z vierohodnostnej funkcie a sú základnými metódami pre testovanie homoskedasticity. Nájdeme ich často popísané v štatistickej a ekonometrickej literatúre (viď napr. [2], [3] alebo [5]). Jedná sa o asymptoticky ekvivalentné testy.

Nešpecifické testy pre heteroskedasticitu

V niektorých prípadoch forma heteroskedasticity nie je známa, ale vieme, že rozptyl chýb monotónne súvisí s veľkosťou nejakej známej exogénnej premennej z . V tomto prípade dva najčastejšie používané testy sú *Goldfeldov–Quandtov test* a *vrcholový („peak“) test*. V prípade domnienky, že heteroskedasticita má tvar $\sigma_i^2 = h(\mathbf{z}_i'\boldsymbol{\alpha})$, kde $h(\cdot)$ je všeobecná funkcia nezávislá na i , môže byť vhodným *Breuschov–Paganov test*. Pokiaľ nemáme žiadnu apriórnu znalosť okrem toho, že heteroskedastické rozptyly $\sigma_i^2 = E(e_i^2)$ sú rovnako ohraňované, použijeme *Whiteov test*.

Goldfeldov–Quandtov test

Jedná sa o veľmi populárny test pre detekciu heteroskedasticity, ktorá monotónne súvisí s exogénnymi premennými a jeho kroky sú nasledovné:

- (1) Zoradíme pozorovania podľa hodnôt premennej z .
- (2) Vyberieme p stredových pozorovaní, ktoré následne vynecháme. Čím viac pozorovaní vynecháme, tým bude vypovedajúca schopnosť testu menšia, takže potrebujeme podmienku, aby počet pozorovaní bol dostatočne veľký, povedzme $(T-p)/2 > K$.

- (3) Prevedieme dve separátne regresie (pomocou *OLS*) na prvú a druhú polovicu zoradených pozorovaní.
- (4) Vypočítame reziduálne súčty štvorcov pre obe skupiny pozorovaní, ktoré označíme SSE_1 (reziduálny súčet štvorcov pre skupinu pozorovaní s malými rozptylmi) a SSE_2 (reziduálny súčet štvorcov pre skupinu pozorovaní s veľkými rozptylmi) a vytvoríme štatistiku

$$F = SSE_2/SSE_1, \quad (3.63)$$

ktorá má za platnosti nulovej hypotézy homoskedasticity F –rozdelenie o $(T - p - 2K)/2$ a $(T - p - 2K)/2$ stupňoch voľnosti.

Je treba poznamenať, že sila Goldfeldovho–Quandtovho testu závisí na p , pričom táto závislosť sa nedá triviálne popísať. Ako sme už spomínali, vynechanie príliš veľkého počtu prostredných pozorovaní vedie k menšej sile testu, ale na druhej strane ponechanie prostredných pozorovaní spôsobí, že rozptyly reziduí sa k sebe približia. Takže voľba p je subjektívna. Navyše sila testu taktiež závisí na relatívnom rozložení \mathbf{z}_i k jeho strednej hodnote. Čím väčší je podiel smerodajnej odchýlky k strednej hodnote \mathbf{z}_i , tým väčšia je sila testu. Ďalej v prípade $p = 0$, Goldfeld–Quandt test je asymptoticky ekvivalentný štatistike založenej na vierohodnostnom pomere (3.53) pre testovanie skupinovej heteroskedasticity s dvoma skupinami. Konečne, podobne ako u testovej štatistiky založenej na vierohodnostnom pomere, predpokladáme, že chybový člen e_i je normálny. Preto Goldfeld–Quandt test je test vhodný aj na malé výbery. Pokiaľ však normalita nie je udržateľná, tak sa musíme spoľahnúť na asymptotické výsledky, prípadne použiť nejaký neparametrický test.

Vrcholový (Peak) test

„Peak“ test sa usiluje o formalizáciu heuristickej grafickej procedúry, ktorá môže byť použitá v prípade, že rozptyl chýb monotónne súvisí s nejakou exogénnou premennou. Jedná sa o neparametrický test, takže nezávisí na normalite porúch. V nasledujúcej časti budeme predpokladať, že rozptyl chýb je monotónne rastúci s premennou z (ale nie je problém procedúru upraviť pre prípad, že rozptyl chýb monotónne klesá s premennou z). „Peak“ test v jednotlivých krokoch je nasledovný:

1. Odhadneme lineárny regresný model pomocou metódy najmenších štvorcov.
2. Nech \hat{e}_i označuje i –te reziduum zodpovedajúce i –temu pozorovaniu z , z_i .
3. Definujme „peak“ (vrchol) ako j –te pozorovanie, kde $|\hat{e}_j| \geq |\hat{e}_i|$ pre všetky $i < j$, pričom prvé reziduum \hat{e}_1 nie je „peak“.
4. K overeniu, či počet takýchto vrcholov (resp. „peakov“) indikuje heteroskedasticitu, sa potrebujeme pozrieť na tabuľku, ktorú uverejnili Goldfeld a Quandt vo svojej kine [8] a kvôli tomu, aby čitateľ tejto práce bol schopný tento test použiť bez ďalšieho pátrania v literatúre, aspoň časť z nej uvádzam v tabuľke 1. Táto tabuľka obsahuje kumulatívne distribúcie počtu peakov za predpokladu nulovej hypotézy homoskedasticity:

Počet peakov $\leq x$											
T	x=0	x=1	x=2	x=3	x=4	x=5	x=6	x=7	x=8	x=9	x=10
5	0,2000	0,6167	0,9083	0,9917	1,0000						
10	0,1000	0,3829	0,7061	0,9055	0,9797	0,9971	0,9997	1,0000			
15	0,0667	0,2834	0,5833	0,8211	0,9433	0,9866	0,9976	0,9997	1,0000		
20	0,0500	0,2274	0,5022	0,7530	0,9056	0,9720	0,9935	0,9988	0,9998	1,0000	
25	0,4000	0,1910	0,4441	0,6979	0,8705	0,9559	0,9879	0,9973	0,9995	0,9999	1,0000
30	0,0333	0,1654	0,4001	0,6525	0,8386	0,9395	0,9815	0,9953	0,9990	0,9998	1,0000
35	0,0286	0,1462	0,3654	0,6144	0,8098	0,9234	0,9745	0,9929	0,9984	0,9997	0,9999
40	0,0250	0,1313	0,3373	0,5818	0,7837	0,9078	0,9674	0,9903	0,9975	0,9995	0,9999
45	0,0222	0,1194	0,3138	0,5536	0,7600	0,8930	0,9601	0,9874	0,9966	0,9992	0,9998
50	0,0200	0,1096	0,2940	0,5288	0,7383	0,8788	0,9530	0,9844	0,9956	0,9989	0,9998
55	0,0182	0,1014	0,2769	0,5068	0,7184	0,8653	0,9456	0,9813	0,9944	0,9986	0,9997
60	0,0167	0,0944	0,2620	0,4871	0,7001	0,8524	0,9384	0,9780	0,9932	0,9982	0,9997

Tabuľka 3.1: Distribučné funkcie rozdelenia peakov za platnosti homoskedasticity

Ak sú chyby e_i heteroskedastické, počet pozorovaných peakov bude veľký. V opačnom prípade nezamietame homoskedasticitu. Treba poznamenať, že tento test nepredpokladá normalitu, jedná sa o asymptotický test, ktorý je platný iba na veľkých výberoch, pretože *OLS*-reziduá nie sú nezávislé na konečných výberoch, pričom Goldfeldova-Quandtova tabuľka predpokladá nezávislé pozorovania. Autori doporučili používať tento test pre rozsah výberu väčší ako 15.

Breusch–Paganov test

Breusch a Pagan predpokladali širšiu triedu heteroskedasticity, definovanú ako

$$\sigma_i^2 = h(\mathbf{z}_i' \boldsymbol{\alpha}), \quad (3.64)$$

kde $h(\cdot)$ je všeobecná funkcia nezávislá na i a $\mathbf{z}_i' \boldsymbol{\alpha}$ je definované rovnako ako v sekcii 3.2. Uvažujme rovnicu

$$\frac{\hat{e}_i^2}{\bar{\sigma}^2} = \mathbf{z}_i' \boldsymbol{\alpha} + \nu_i, \quad (3.65)$$

kde \hat{e}_i^2 je i -te *OLS*-reziduum a $\bar{\sigma}^2 = \sum_{i=1}^T \hat{e}_i^2 / T$. *OLS*-odhad $\hat{\boldsymbol{\alpha}}$, ktorý dostaneme z (3.65) je konzistentný, ale pokiaľ nepoznáme presnú formu $h(\cdot)$, *FGLS* nemôžeme použiť. Ale ak nepoznáme konkrétnu formu $h(\cdot)$, tento test pre heteroskedasticitu môžeme stále požiť, pokiaľ sú chyby normálne. Nech RSS označuje regresný súčet štvorcov získaný ako *OLS*-odhad z (3.65), t.j. nech $y_i = \hat{e}_i^2 / \bar{\sigma}^2$, $\bar{y}_i = \sum_{i=1}^T y_i / T$ a $\hat{y}_i = \mathbf{z}_i' \boldsymbol{\alpha}$, potom $RSS = \sum_{i=1}^T (\hat{y}_i - \bar{y})^2$. Breusch a Pagan ukazujú, že ak $\boldsymbol{\alpha}_0 = 0$, tak asymptoticky platí:

$$\frac{1}{2} RSS \sim \chi_{p-1}^2. \quad (3.66)$$

Takže pokiaľ je hodnota testovej štatistiky väčšia ako kritická hodnota χ_{p-1}^2 , nulovú hypotézu homoskedasticity zamietame v prospech alternatívnej hypotézy, ktorá predpokladá heteroskedasticitu v tvare $\sigma_i^2 = h(\mathbf{z}_i' \boldsymbol{\alpha})$.

Whiteov test

White rieši prípad, keď nemáme žiadnu ďalšiu informáciu o štruktúre heteroskedasticity, okrem toho, že rozptyly sú heteroskedastické $\sigma_i^2 = E(e_i^2)$ a rovnako ohraničené.

Ako už bolo uvedené, nie je možné odhadnúť každé σ_i^2 , pretože by sme odhadovali T parametrov, pričom máme k dispozícii iba T pozorovaní. To však nevyklučuje konzistentný odhad kovariančnej matice OLS -odhadu $\hat{\beta}$, a to $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}\mathbf{\Omega})(\mathbf{X}'\mathbf{X})^{-1}$ keď $\mathbf{E}\mathbf{e}\mathbf{e}' = \mathbf{\Omega} = \text{diag}\{\sigma_1^2, \sigma_2^2, \dots, \sigma_T^2\}$ je GLS -štruktúra chýb. Za všeobecných podmienok, $\hat{\mathbf{V}} = \sum_{i=1}^T \hat{e}_i^2 \mathbf{x}_i \mathbf{x}_i'$, kde \mathbf{x}_i' označuje i -ty riadok matice \mathbf{X} a $\hat{e}_i^2 = (y_i - \mathbf{x}_i' \hat{\beta})^2$, je konzistentný odhad $(\mathbf{X}'\mathbf{\Omega}\mathbf{X})$. Takže kovariančnú maticu OLS -odhadu môžeme konzistentne odhadnúť ako $(\mathbf{X}'\mathbf{X})^{-1} \hat{\mathbf{V}} (\mathbf{X}'\mathbf{X})^{-1}$. Závěry o β sú ešte stále možné prostredníctvom OLS -odhadu $\hat{\beta}$, aj keď špecifickú štruktúru $\mathbf{\Omega}$ nepoznáme, pretože $\hat{\beta}$, má asymptoticky normálne rozdelenie. Všeobecnejšie, White ukazuje, že test všeobecnej lineárnej hypotézy $H_0 : \mathbf{R}\beta = \mathbf{r}$ proti $H_1 : \mathbf{R}\beta \neq \mathbf{r}$, môže byť prevedený a poznamenáva, že za platnosti nulovej hypotézy H_0 asymptoticky platí:

$$(\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \hat{\mathbf{V}} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \chi_J^2; \quad (3.67)$$

kde J označuje počet reštrikcií.

Whiteov test pre heteroskedasticitu, ktorý pozostáva z porovnávania prvkov $\hat{\mathbf{V}}$ a $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})$, čím zistíme, či bežný OLS -odhad kovariančnej matice $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ je alebo nie je konzistentným odhadom. Veľké rozdiely medzi $\hat{\mathbf{V}}$ a $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})$ podporujú tvrdenie o heteroskedasticite, kým malé rozdiely naznačujú homoskedasticitu.

Whiteov test je veľmi používaný (najmä v ekonometrickom software), avšak je pomerne nekonštruktívny, v prípade zamietnutia homoskedasticity vôbec nenaznačí ďalší postup. Pre názornosť jeho princíp ešte ukážeme na príklade:

Chceme testovať homoskedasticitu proti alternatívnej hypotéze heteroskedasticity. Máme napríklad model

$$y_i = \beta_i + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i, \quad i = 1, 2, \dots, T.$$

Whiteov test vytvorí pomocný model, ktorý je lineárnou regresiou štvorcov OLS -reziduí až na konštantu:

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i2}^2 + \alpha_5 x_{i3}^2 + \alpha_6 x_{i2} x_{i3} + u_i.$$

Chceme zistiť, či sa rozptyl pôvodných chýb systematicky mení v závislosti na regresoroch pôvodného modelu. Pomocný model musí obsahovať konštantný člen aj keby nebol v pôvodnom modele, pretože štvorce OLS -reziduí majú kladnú strednú hodnotu. Pomocou tohoto pomocného modelu prevedieme χ^2 -test. Pre χ^2 -test stačí nájsť v pomocnom modele koeficient determinácie R^2 :

$$R^2 = \left(\sum_{i=1}^T (\hat{e}_i - \bar{e})^2 \right) / \left(\sum_{i=1}^T (e_i - \bar{e})^2 \right)$$

a príslušný kritický obor nulovej hypotézy na hladine α je potom

$$(T - 6) \cdot R^2 \geq \chi_{1-\alpha}^2(5),$$

pričom máme 6 regresorov a 5 obmedzení.

3.2.5 Chowov test

V mnohých prípadoch, vzniká potreba zistiť, či chovanie alebo výkon v jednej skupine je odlišný od ostatných skupín alebo či v čase nastali nejaké štrukturálne zmeny. Uvažujme model

$$\begin{aligned} \mathbf{y}_1 &= X_1 \boldsymbol{\beta}_1 + \mathbf{e}_1, \\ \mathbf{y}_2 &= X_2 \boldsymbol{\beta}_2 + \mathbf{e}_2, \end{aligned} \quad (3.68)$$

kde \mathbf{y}_i a X_i sú $T_i \times 1$ a $T_i \times K$ matice pozorovaní závislých a nezávislých premenných, $\boldsymbol{\beta}_i$ sú pravdepodobne odlišné vektorové koeficienty typu $K \times 1$ a \mathbf{e}_i sú chybové vektory typu $T_i \times 1$ s centrovaným normálnym rozdelením

$$\begin{aligned} \mathbf{e}_1 &\sim N(\mathbf{0}, \sigma_1^2 \mathbf{I}_{T_1}), \\ \mathbf{e}_2 &\sim N(\mathbf{0}, \sigma_2^2 \mathbf{I}_{T_2}), \end{aligned} \quad (3.69)$$

kde σ_1^2 a σ_2^2 nemusia byť rovnaké. V štrukturálnej analýze chceme testovať

$$\begin{aligned} H_0 : \boldsymbol{\beta}_1 &= \boldsymbol{\beta}_2, \\ H_1 : \boldsymbol{\beta}_1 &\neq \boldsymbol{\beta}_2. \end{aligned}$$

V prípade, že $\sigma_1^2 = \sigma_2^2$, jedná sa o klasický F -test s testovou štatistikou u s F -rozdelením o K a $T_1 + T_2 - 2K$ stupňoch voľnosti:

$$u = \frac{(SSE_r - SSE_u)/K}{SSE_u/(T_1 + T_2 - 2K)}, \quad (3.70)$$

kde SSE_u je súčet štvorcových chýb (reziduálny súčet štvorcov), ktorý získame aplikáciou bežnej metódy najmenších štvorcov na systém rovníc

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

, t.j. $SSE_u = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ a SSE_r je súčet štvorcových chýb zúženého modelu:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \boldsymbol{\beta}_\cdot + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} = X_\cdot \boldsymbol{\beta}_\cdot + \mathbf{e}_\cdot,$$

kde $\boldsymbol{\beta}_\cdot$ je vektor $K \times 1$, pričom $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ a $SSE_r = (\mathbf{y} - X_\cdot \hat{\boldsymbol{\beta}}_\cdot)'(\mathbf{y} - X_\cdot \hat{\boldsymbol{\beta}}_\cdot)$, $\hat{\boldsymbol{\beta}}_\cdot = (X_\cdot' X_\cdot)^{-1} X_\cdot' \mathbf{y}$.

Chowov test býva najčastejšie využívaný v analýze časových radov k odhaleniu štrukturálnych skokov.

Jayatissaov test

Jayatissa v roku 1977 skonštruoval vhodnú testovú štatistiku, pre Chowov test za heteroskedasticity, ktorá je použiteľná na malých výberoch. Uvažujme opäť model

(3.68). Nech odhad parametru β je $\hat{\beta}_i = (X_i'X_i)^{-1}X_i'\mathbf{y}_i$ a nech ďalej $\hat{\mathbf{e}}_i = M_i\mathbf{y}_i$, kde $M_i = \mathbf{I}_{T_i} - X_i(X_i'X_i)^{-1}X_i' = Z_i'Z_i$ s $Z_i'X_i = 0$ a $Z_i'Z_i = \mathbf{I}_{(T_i-K)}$. Z_i je ortonormálna matica typu $[T_i \times (T_i - K)]$ so stĺpcami, ktoré sú vlastnými vektormi matice M_i . Jayitissov test sa skladá z nasledujúcich krokov:

- (1) Spočítajme $\mathbf{d} = (\hat{\beta})_1 - (\hat{\beta})_2$.
- (2) Spočítajme $Z_i, i = 1, 2$.
- (3) Spočítajme $r \leq \min[(T_1 - K)/K, (T_2 - K)/K]$.
- (4) Spočítajme vektory $Z_i'\hat{\mathbf{e}}_i, i = 1, 2$. Vyberme nejakých r subvektorov z $Z_1'\hat{\mathbf{e}}_1$ a tiež nejakých r subvektorov z $Z_2'\hat{\mathbf{e}}_2$ dimenzie $K \times 1$ a označme ich zaradom $\mathbf{e}_{1(1)}^*, \mathbf{e}_{1(2)}^*, \dots, \mathbf{e}_{1(r)}^*$ a $\mathbf{e}_{2(1)}^*, \mathbf{e}_{2(2)}^*, \dots, \mathbf{e}_{2(r)}^*$.
- (5) Spočítajme štvorcovú maticu Q_i typu $K \times K$, pre ktorú platí $Q_i'Q_i = (X_i'X_i)^{-1}$.
- (6) Spočítajme $\boldsymbol{\eta}_j = Q_1'\mathbf{e}_{1((j))}^* + Q_2'\mathbf{e}_{2((j))}^*, j = 1, 2, \dots, r$.
- (7) Spočítajme

$$S = \frac{1}{r} \sum_{j=1}^r \boldsymbol{\eta}_j \boldsymbol{\eta}_j'.$$

Potom za nulovej hypotézy $H_0 : \beta_1 = \beta_2$, kde $\sigma_1^2 \neq \sigma_2^2$ a $r \geq K$, testová štatistika

$$J = \frac{\mathbf{d}'S^{-1}\mathbf{d}}{r} \cdot \frac{(r - K + 1)}{K} \quad (3.71)$$

má F –rozdelenie o K a $(r - K + 1)$ stupňoch voľnosti.

3.3 Autokorelácia

k nej dôjsť aj K porušeniu predpokladu nekorelovaných reziduí dochádza pomerne často u časových radov, ale môže u prierezových dát usporiadaných napríklad podľa hodnôt niektorého regresoru. K autokorelácii dochádza z rôznych príčin. Častým dôvodom je, že v modele chýbajú nejaké regresory, ktorých pozorovania sú autokorelované, a tak sa autokorelovanosť preniesie do reziduí, prípadne sme nedostatočne špecifikovali dynamiku modelu, t.j. do modelu sme nezaradili nejaké omeškané premenné (závislú alebo nezávislú), ktoré sme mali zaradiť medzi regresory. Ďalším dôvodom autokorelácie reziduí môže byť napríklad aj nevhodne zvolený tvar použitej regresnej krivky, atď. Ignorácia autokorelácií vedie k neeficientnému odhadu modelu a k horšej schopnosti predpovedať budúce hodnoty.

V prípade, že nemôžeme kvôli vzájomnej korelovanosti použiť klasickú regresiu, ponúka sa nám ako alternatíva Boxova–Jenkinsova metodológia. Box a Jenkins v [4] uviedli základné princípy tohoto prístupu. Základným rozdielom oproti klasickej regresii je fakt, že stavebným prvkom v prístupe Boxa a Jenkinsa je reziduálna zložka, ktorá je dokonca obvykle tvorená korelovanými náhodnými veličinami, nakoľko Box a Jenkins vychádzali z korelačnej analýzy. Napriek tomu, že tento prístup vychádza z reziduálnej zložky, nestrácame možnosť modelovať trend a ani sezónnosť. Boxov–Jenkinsov model však nie je možné navrhnúť hypoteticky, na základe nejakej teórie bez dátovej analýzy, nakoľko dosiahnutá flexibilita je výsledkom vývoja modelu z nameraných dát.

Vývoj Boxových–Jenkinsových modelov sa dá rozdeliť do troch krokov: identifikácia modelu, odhad parametrov a overenie modelu. V prípade, že sa model pri overovaní nepotvrdí, musíme všetky tri kroky zopakovať. Je nutné poznamenať, že Boxova–Jenkinsova metodológia je uplatniteľná len v prípade, že k dispozícii máme dostatok pozorovaní, v literatúre sa ako minimálny počet pozorovaní uvádza zvyčajne 50. Potreba pomerne veľkého počtu pozorovaní nie je jediným obmedzením a nevýhodou použitia týchto metód. Podstatnou nepríjemnosťou, ktorú prináša fakt, že model vychádza iba z napozorovaných dát, je obtiažny spôsob interpretácie výsledných modelov. Vhodné použitie tohto prístupu dáva však dobré predpovede, ktoré obstoja ako argument voči problematickej interpretácii modelu. Ďalším, ale väčšinou pomerne jednoducho odstrániteľným obmedzením pre tento prístup, je požiadavka na stacionaritu časového radu (prípadne usporiadaných prierezových dát), ktorý zvyčajne môžeme odstrániť pomocou nejakej vhodnej transformácie.

3.3.1 Autokorelačné vlastnosti časových radov

V tejto časti budú popísané základné definície nutné pre pochopenie Boxovej–Jenkinsovej metodológie a pre výstavbu modelov založených na reziduálnej zložke. Ako už bolo spomenuté, pri výstavbe Boxových–Jenkinsových modelov je dôležitá stacionarita časového radu. Jedná sa v podstate o určité stabilné chovanie časového radu/procesu v čase. V množstve štatistickej literatúry (napr. v [2]) sú uvádzané definície *striktnej* a *slabej stacionarity*. V tomto prípade si však vystačíme so *sla-*

bou *stacionaritou*, o ktorej budeme hovoriť iba ako o *stacionarite*, a ktorej definíciu uvádzame:

Definícia 3.6 Náhodný proces $\{y_t, t \in T\}$ nazývame *lineárny proces*, pokiaľ je možné zapísať ho v tvare nekonečnej sumy

$$y_t = \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \dots, \quad (3.72)$$

kde ϵ_t je biely šum s rozptylom σ_ϵ^2 a ψ_j sú parametre. Ekvivalentne môžeme lineárny proces zapísať pomocou tzv. operátora spätného posunutia B ako

$$y_t = \psi(B)\epsilon_t, \quad (3.73)$$

kde $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \dots = 1 + \sum_{j=1}^{\infty} \psi_j B^j$ a B je definovaný pomocou

$$By_t = y_{t-1} \quad (3.74)$$

a ktorý môžeme aplikovať viacnásobne ako $B^j y_t = y_{t-j}$.

V [4] je ukázané, že konvergencia $\psi(B)$ podľa kvadratického stredy je postačujúcou podmienkou pre existenciu lineárneho procesu (3.72). Ak je $|B| \leq 1$, potom lineárny proces (3.72) existuje, je centrováný ($E(y_t) = 0$) a stacionárny (viď definícia (3.7)). V prípade, že lineárny proces (3.72) ide prepísať do tvaru

$$y_t = \pi_1 y_{t-1} + \pi_2 y_{t-2} + \dots + \epsilon_t, \quad (3.75)$$

hovoríme, že lineárny proces je *invertibilý*, čo môžeme tiež zapísať pomocou operátora spätného posunutia ako $\pi(B)y_t = \epsilon_t$. Opäť v [4] nájdeme, že konvergencia $\pi(B)$ podľa kvadratického stredy je postačujúcou podmienkou pre invertibilitu lineárneho procesu, t.j. $\pi(B)$ konverguje pre $|B| \leq 1$.

Definícia 3.7 Náhodný proces $\{y_t, t \in T\}$ s konečnými druhými momentami nazývame *slabo stacionárny* (ďalej len *stacionárny*) práve vtedy, keď má konštantnú strednú hodnotu, $E(y_t) = \mu$, pre všetky $t \in T$ a ak jeho autokovariančná funkcia $\text{cov}(y_t, y_{t+k}) = E(y_t - E(y_t))(y_{t+k} - E(y_{t+k}))$ závisí iba na rozdieli svojich argumentov $t+k-t=k$. V tomto prípade označíme autokovariančnú funkciu $\gamma_k = \text{cov}(y_t, y_{t+k})$.

Z definície je zrejmé, že stacionárny proces má konštantné rozptyly a kovariančná štruktúra druhého rádu je invariantná voči posunom v čase. Kovariancia sa zvykne často normovať, čím dostávame *korelačný koeficient*:

$$\text{cor}(y_t, y_{t+k}) = \frac{\text{cov}(y_t, y_{t+k})}{\sqrt{\text{var}(y_t)\text{var}(y_{t+k})}} \quad (3.76)$$

a v prípade stacionárneho procesu môžeme písať $\text{cor}(y_t, y_{t+k}) = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma_y^2} = \rho_k$, pre $k = \dots, -1, 0, 1, \dots$, kde $\sigma_y^2 = \gamma_0 = \text{var}(y_t)$ je rozptyl radu $\{y_t\}$. Ďalším pojmom, s ktorým budeme ďalej pracovať je *parciálna autokorelačná funkcia* ρ_{kk} . Jedná sa

o parciálny korelačný koeficient y_t a y_{t+k} pri pevných hodnotách $y_{t+1}, \dots, y_{t+k-1}$. Výpočet ρ_{kk} prebieha podľa [4] nasledujúcim spôsobom:

$$\rho_{kk} = \frac{|\mathbf{P}_k^*|}{|\mathbf{P}_k|}, \quad (3.77)$$

t.j. parciálna autokorelačná funkcia je podielom determinantov matice \mathbf{P}_k^* a matice autokorelácií \mathbf{P}_k , kde

$$\mathbf{P}_k = \begin{bmatrix} 1 & \rho_1 & \dots & \rho_{k-1} \\ \rho_1 & 1 & \dots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \dots & 1 \end{bmatrix}$$

a \mathbf{P}_k^* vznikne predefinovaním posledného stĺpca v \mathbf{P}_k :

$$\mathbf{P}_k^* = \begin{bmatrix} 1 & \rho_1 & \dots & \rho_1 \\ \rho_1 & 1 & \dots & \rho_2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \dots & \rho_k \end{bmatrix}$$

Je dôležité všimnúť si, že autokovariančná funkcia γ_k i autokorelačná funkcia ρ_k sú párne funkcie, takže stačí uvažovať $k \geq 0$. Graf autokorelačnej funkcie ρ_k sa nazýva *korelogram*.

Nakoľko Boxova–Jenkinsova metodológia používa pre výstavbu modelu výhradne len napozorované dáta, s teoretickou autokovarianciou ani autokoreláciou nevystačíme. Uvedme teda ešte odhady strednej hodnoty, autokovariančnej, autokorelačnej a parciálnej autokorelačnej funkcie pre napozorovaný časový rad y_1, \dots, y_n , ktoré sa bežne pri výstavbe Boxových–Jenkinsových modelov používajú.

Pre časovú radu y_1, y_2, \dots, y_n môžeme použiť (a budeme používať) ako *odhad strednej hodnoty* aritmetický priemer napozorovaného radu

$$\bar{y} = \sum_{t=1}^n y_t / n, \quad (3.78)$$

ako *odhad autokovariančnej funkcie*

$$c_k = \sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y}) / n, \quad (3.79)$$

logicky potom *odhad autokorelačnej funkcie* bude

$$r_k = c_k / c_0, \quad (3.80)$$

a konečne *odhad autokorelačnej funkcie*

$$\begin{aligned} r_{11} &= r_1, \\ r_{kk} &= \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_j}, \end{aligned} \quad (3.81)$$

kde

$$r_{kj} = r_{k-1,j} - r_{kk}r_{k-1,k-j}, \quad j = 1, 2, \dots, k-1 \quad (3.82)$$

pričom autokorelačnú a autokovariančnú funkciu je možné odhadovať pre indexy $k = 0, 1, \dots, n-1$ a okrem už spomínaného $n > 50$ sa zvyčajne kvôli použiteľnosti odhadov vyžaduje $k < n/4$. Chovanie autokorelačnej funkcie je v Boxovej–Jenkinsovej metodológii veľmi dôležité, nakoľko identifikuje vhodný model. Najdôležitejšie je určiť existenciu a veľkosť tzv. *identifikačného bodu*, resp. *bodu useknutia*, od ktorého začína byť autokorelačná funkcia nulová. T.j. potrebujeme zistiť, či existuje k_0 také, že pre všetky $k > k_0$ môžeme tvrdiť, že ρ_k je nulová. K určeniu existencie bodu useknutia, resp. k určeniu samotného bodu useknutia k_0 sa väčšinou (za predpokladu normality) používa *Bartletova aproximácia*:

$$\sigma(r_k) = \sqrt{\text{var}(r_k)} \sim \sqrt{\frac{(1 + 2 \sum_{j=1}^{k_0} r_j^2)}{n}}, \quad k > k_0. \quad (3.83)$$

Za predpokladu (asymptotickej) normality r_k s (asymptotickou) strednou hodnotou ρ_k môžeme považovať ρ_k za nulové, ak $|r_k| < 2\sigma(r_k)$. Dôvodom pre tento záver je fakt, že 95% konfidenčný interval pre normálnu náhodnú veličinu s nulovou strednou hodnotou a rozptylom σ^2 je $(-2\sigma, 2\sigma)$.

Obdobne potrebujeme testovať aj nulovosť parciálnej autokorelačnej funkcie ρ_{kk} , t.j. bod useknutia pre ρ_{kk} . Pre tento účel sa používa *Quenouilleova aproximácia* pre smerodajnú odchýlku r_{kk} . Platí, že ak je $\rho_{kk} = 0$ pre $k > k_0$, potom

$$\sigma(r_{kk}) \sim \sqrt{1/n}, \quad k > k_0. \quad (3.84)$$

Takže v prípade, že existuje k_0 také, že pre všetky $k > k_0$ je $r_{kk} < 2\sqrt{1/n}$ môžeme r_{kk} považovať za nulovú pre všetky $k > k_0$.

Definovaný lineárny proces je teoretickým základom pre Boxove–Jenkinsove modely, s ktorými sa ďalej oboznámime. Obecný lineárny proces však nie je v praxi tak významný ako jeho špeciálne prípady.

3.3.2 Autoregresná postupnosť prvého rádu AR(1)

Ako už bolo spomenuté, častým dôvodom pre autokoreláciu chybovej zložky býva vynechanie nejakej premennej, ktorá by do modelu mala patriť. Napr. keď „skutočný“ regresný vzťah je

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 z_t + e_t \quad (3.85)$$

ale miesto toho použijeme vzťah:

$$y_t = \beta_1 + \beta_2 x_{t2} + u_t, \quad (3.86)$$

kde $u_t = \beta_3 z_t + e_t$. Ak je z_t autokorelované, nemá v čase žiaden trend a ak je stochasticky nezávislé na x_{t2} , potom u_t je autokorelované a nezávislé na x_{t2} . S týmto prípadom sa ešte budeme zaoberať, avšak v prípade, že vynechaný regresor je súčasne korelovaný s x_{t2} , odhad (*OLS* odhad ako aj odhad pomocou metód opísaných v tejto

časti) bude vychýlený a dáva nekonzistentné výsledky. Takže vidíme, že určitý zdroj autokorelovanosti chýb je veľmi dôležitý. V prípade, že autokorelácia vzniká vynechaním autokorelovanej premennej, ktorá súčasne súvisí s nejakou použitou vysvetľujúcou premennou, tak by sme sa mali usilovať o zaradenie vynechanej premennej do modelu alebo aspoň v prípade jej vylúčenia zaznamenať potenciálne vychýlenie, ktoré môže vzniknúť na základe vynechania tejto premennej. Ďalej budeme predpokladať, že skúmaná autokorelovaná chybová zložka nesúvisí so žiadnym použitým regresorom, ale miesto toho chyby reprezentujú nezávislé pretrvávajúce šoky, vznikajúce zo zotrvačnosti (často ekonomických) procesov.

Jedným zo spôsobov ako môže byť závislosť chýb znázornená (a v praxi je schopná pokryť veľkú časť prípadov) je lineárny model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (3.87)$$

pričom ale

$$e_t = \rho e_{t-1} + u_t, \quad t = 1, 2, \dots, T \quad (3.88)$$

a pre parameter ρ predpokladáme, že $|\rho| < 1$, u_t sú *i.i.d.* s nulovou strednou hodnotou a s rozptylom σ_u^2 (t.j. u_t je *biely šum*) a stacionárny stochastický proces generujúci e_t začal v nekonečnej minulosti. Takýto lineárny proces $\{e_t\}$ nazývame autoregresný proces 1. rádu alebo AR(1). V autoregresnom vzťahu (3.88) je veľmi podstatné znamienko parametru ρ , pretože v prípade kladného ρ hovoríme o tzv. *pozitívnej autokorelovanosti* s riedkymi zmenami znamienka reziduí e_t , a naopak v prípade záporného ρ o *negatívnej autokorelovanosti* s častými zmenami znamienka e_t . Postupnou spätnou substitúciou dostaneme:

$$e_t = \sum_{i=0}^{\infty} \rho^i u_{t-i}. \quad (3.89)$$

Vidíme, že $\{e_t\}$ je špeciálny prípad lineárneho procesu, ktorý sme už zadefinovali a z predpokladu na u_t ďalej plyní, že sa jedná o centrováný proces:

$$\mathbb{E}(e_t) = \sum_{i=0}^{\infty} \rho^i \mathbb{E}(u_{t-i}) = 0 \quad (3.90)$$

s konštantným rozptylom:

$$\begin{aligned} \mathbb{E}(e_t^2) &= \mathbb{E}(u_t^2) + \rho^2 \mathbb{E}(u_t^2) + \rho^4 \mathbb{E}(u_t^2) + \dots \\ &= \sigma_u^2 / (1 - \rho^2) = \sigma_e^2. \end{aligned} \quad (3.91)$$

Kovariancia medzi e_t a e_{t-i} je

$$\begin{aligned} \mathbb{E}(e_t e_{t-i}) &= \mathbb{E}([u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \dots] \times [u_{t-i} + \rho u_{t-i-1} + \rho^2 u_{t-i-2} + \dots]) \\ &= \rho^i \sigma_e^2. \end{aligned} \quad (3.92)$$

Takže variančná matica vektoru reziduí \mathbf{e} má tvar:

$$\mathbf{E}\mathbf{e}\mathbf{e}' = \sigma_u^2 \mathbf{\Omega} = \sigma_u^2 \frac{1}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & & \rho^{T-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \rho^{T-2} & \rho^{T-3} & & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \dots & \rho & 1 \end{bmatrix}.$$

Pretože variančná matica chýb je *nesférická*, t.j. nie je skalárnym násobkom jednotkovej matice, metóda najmenších štvorcov poskytne síce nestranný, ale ako už bolo spomínané, pomerne neúčinný odhad vzhľadom k zobecnenej metóde najmenších štvorcov. Odhad zobecnenu metódou najmenších štvorcov je:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}, \quad (3.93)$$

kde

$$\mathbf{\Omega}^{-1} = \begin{bmatrix} 1 & -\rho & & & \\ -\rho & (1 + \rho^2) & -\rho & & 0 \\ & -\rho & (1 + \rho^2) & & \\ & & & \ddots & -\rho \\ 0 & & & & (1 + \rho^2) & -\rho \\ & & & & -\rho & 1 \end{bmatrix}.$$

V prípade, že $\mathbf{\Omega}$ je známa, zobecnenú metódu najmenších štvorcov môžeme transformovať nasledujúcim spôsobom:

$$\mathbf{P}'\mathbf{y} = \mathbf{P}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}'\mathbf{e}, \quad (3.94)$$

kde

$$\mathbf{P}' = \begin{bmatrix} (1 - \rho^2)^{1/2} & & & & \\ & -\rho & 1 & & 0 \\ & & -\rho & 1 & \\ & 0 & & \ddots & \\ & & & & -\rho & 1 \end{bmatrix}$$

je Praisova-Winstenova transformačná matica typu $T \times T$. Chyby transformovaného modelu sú *i.i.d.* a tak metódou najmenších štvorcov aplikovanou na (3.94) získame zobecnený odhad najmenších štvorcov parametru $\boldsymbol{\beta}$. Ak je ρ známe, zobecnenú metódu najmenších štvorcov ľahko spočítame a za normality pri malom výbere získame najlepší nestranný odhad, pri veľkom výbere získame asymptoticky eficientný odhad. Častejšie je však ρ neznáme a musíme vypočítať konzistentný odhad $\hat{\rho}$. V zobecnenej metóde najmenších štvorcov potom použijeme miesto ρ jeho konzistentný odhad $\hat{\rho}$ a tým získame *FGLS* odhad.

Maximálne vierohodný odhad v AR(1)

Maximálne vierohodné odhady $\boldsymbol{\beta}$, ρ a σ_u^2 získame ako obvykle pomocou logaritmickkej vierohodnostnej funkcie, ktorá má v tomto modele tvar:

$$\begin{aligned}
 \ln L(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}, \sigma_u^2, \rho) &= -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \ln |\sigma_u^2 \boldsymbol{\Omega}| - \frac{1}{2\sigma_u^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
 &= -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma_u^2) + \frac{1}{2} \ln(1 - \rho^2) \\
 &\quad - \frac{1}{2\sigma_u^2} (\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta})' (\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}),
 \end{aligned} \tag{3.95}$$

kde $\mathbf{y}^* = \mathbf{P}'\mathbf{y}$, $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$ a $|\sigma_u^2 \boldsymbol{\Omega}| = (\sigma_u^2)^T / (1 - \rho^2)$.

Maximalizáciou vzhľadom k $\boldsymbol{\beta}$ a σ_u^2 dostaneme

$$\boldsymbol{\beta}(\rho) = (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{y}^* = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{y} \tag{3.96}$$

a

$$\sigma_u^2(\rho) = \frac{1}{T} [\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}(\rho)]' [\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}(\rho)]. \tag{3.97}$$

Spätnou substitúciou (3.96) a (3.97) do virohodnostnej funkcie dostaneme koncentrovanú virohodnostnú funkciu

$$\begin{aligned}
 L^*(\rho; \mathbf{y}, \mathbf{X}) &= -\frac{T}{2} (\ln(2\pi) + 1) + \frac{1}{2} \ln(1 - \rho^2) - \frac{T}{2} \ln \sigma_u^2(\rho) = \\
 &= \frac{T}{2} (\ln(2\pi) + 1) - \frac{T}{2} \ln \left[\frac{\sigma_u^2(\rho)}{(1 - \rho^2)^{1/T}} \right].
 \end{aligned}$$

Takže globálne maximum logaritmickej virohodnostnej funkcie je rovné globálnemu minimu funkcie $\sigma_u^2(\rho)/(1 - \rho^2)^{1/T}$. Maximálne virohodné odhady $\boldsymbol{\beta}$, ρ a σ_u^2 sú ako obvykle konzistentné a asymptoticky efficientné.

Cochraneova-Orcuttova iteračná procedúra

Praisovu – Winstenovu transformačnú maticu nie je možné použiť v prípade, že $\boldsymbol{\Omega}$ je neznáma. Cochrane a Orcutt navrhli použiť transformáciu, ktorá vynechá prvé pozorovanie:

$$\mathbf{G} = \begin{bmatrix} -\rho & 1 & & \\ & -\rho & 1 & 0 \\ & & \ddots & \ddots \\ 0 & & \ddots & \ddots \\ & & & -\rho & 1 \end{bmatrix}_{(T-1) \times T}.$$

Cochraneov–Orcuttov transformačný systém je

$$y_t - \rho y_{t-1} = \beta_1(x_{t1} - \rho x_{t-1,1}) + \dots + \beta_K(x_{tK} - \rho x_{t-1,K}) + e_t - \rho e_{t-1},$$

kde $t = 1, 2, \dots, T$ t.j. v maticovom zápise

$$\mathbf{G}\mathbf{y} = \mathbf{G}\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{e}, \tag{3.98}$$

kde \mathbf{y} reprezentuje vektor $T \times 1$ pozorovaní závislej premennej, \mathbf{X} je konštrukčná matica $T \times K$ a \mathbf{e} je vektor chýb $T \times 1$. Odhad

$$\tilde{\boldsymbol{\beta}}_{\mathbf{G}} = (\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}'\mathbf{G}\mathbf{y} \quad (3.99)$$

je maximálne vierohodný odhad $\boldsymbol{\beta}$ pre známe ρ podmienené daným y_t .

Pre prípad, že ρ je neznáme Cochrane a Orcutt navrhli iteračnú procedúru pre odhad $\boldsymbol{\beta}$. Nech

$$\hat{e}_t = y_t - (x_{t1}\hat{\beta}_1 + \dots + x_{tK}\hat{\beta}_K)$$

označujú najmenšie reziduálne štvorce pre uvažovaný lineárny model (3.87). Konzistentný odhad ρ dostaneme ako $\hat{\rho} = \sum_{t=2}^T \hat{e}_t \hat{e}_{t-1} / \sum_{t=2}^T \hat{e}_{t-1}^2$. A tento odhad následne použijeme do odhadu $\boldsymbol{\beta}$ ako

$$\tilde{\tilde{\boldsymbol{\beta}}}_{\mathbf{G}} = (\mathbf{X}'\hat{\mathbf{G}}'\hat{\mathbf{G}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{G}}'\hat{\mathbf{G}}\mathbf{y}, \quad (3.100)$$

kde $\hat{\mathbf{G}}$ označuje maticu \mathbf{G} v ktorej je miesto ρ použitý odhad $\hat{\rho}$. Ďalej použitím $\tilde{\tilde{\boldsymbol{\beta}}}_{\mathbf{G}}$ získame nový odhad ρ ako $\hat{\hat{\rho}} = \sum_{t=2}^T \hat{e}_t \hat{\hat{e}}_{t-1} / \sum_{t=2}^T \hat{\hat{e}}_{t-1}^2$, kde $\hat{\hat{e}}_t = y_t - (x_{t1}\hat{\hat{\beta}}_1 + \dots + x_{tK}\hat{\hat{\beta}}_K)$ a $\hat{\hat{\beta}}_i$ pre $i = 1, 2, \dots, K$ je prvok z $\tilde{\tilde{\boldsymbol{\beta}}}_{\mathbf{G}}$. Takto pokračujeme až kým dve po sebe idúce iterácie nedajú výsledky dostatočne blízko pri sebe.

Hildrethova-Luova vyhľadávacia metóda

Hilderth a Lu navrhli aby sa po iniciálnom odhade nepokračovalo iteračnou procedúrou, ale aby sa interval $(-1, 1)$ rozdelil po bodoch ρ_i , $i = 1, 2, \dots, n$. Potom pre každé ρ_i spočítame zodpovedajúci súčet štvorcových chýb $SSE(\rho_i)$ pre Cochraneovu–Orcuttovu transformáciu. Ako výsledný odhad ρ označíme ρ_i , ktoré minimalizuje $SSE(\rho_i)$ pomocou ktorého následne dopočítame odhady $\beta_1, \beta_2, \dots, \beta_K$.

Praisova–Winstenova, Cochraneova–Orcuttova procedúra a Hilderthova–Luova procedúra sú asymptoticky ekvivalentné.

Dvojkrokové metódy

Uvažujme dva rôzne zobecnené odhady metódou najmenších štvorcov a vychádzajúce z Paris-Winstenovej a Cochrane-Orcuttovej transformácie. Keď ρ nahradíme konzistentným odhadom $\hat{\rho}$ dostaneme *FGLS* odhad. Je viac možností ako konzistentne odhadnúť ρ .

1. Cochraneova–Orcuttova metóda: chyby e_t nahradíme v rovnosti

$$e_t = \rho e_{t-1} + u_t, \quad t = 2, 3, \dots, T$$

odhadom metódou najmenších štvorcov \hat{e}_t a aplikujme metódu najmenších štvorcov na autoregresnú rovnosť. Odhad metódou najmenších štvorcov je

$$\hat{\hat{\rho}} = \frac{\sum_{t=2}^T \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^T \hat{e}_{t-1}^2}.$$

Tento odhad je konzistentný, ale na malých výberoch je vychýlený.

2. Theilov odhad:

$$\hat{\rho} = \frac{\frac{1}{T-1} \sum_{t=2}^T \hat{e}_t \hat{e}_{t-1}}{\frac{1}{T-K} \sum_{t=1}^T \hat{e}_t^2} = \frac{\sum_{t=2}^T \hat{e}_t \hat{e}_{t-1}}{(T-1) \hat{\sigma}^2}.$$

3. Durbinov odhad: nech

$$y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + \rho e_t,$$

kde

$$e_t = \rho e_{t-1} + u_t.$$

a dostaneme transformáciu modelu

$$\begin{aligned} y_t &= \rho y_{t-1} + (1 - \rho) \beta_1 + \beta_2 x_{t2} - \beta_2 \rho x_{t-1,2} + \dots \\ &\quad + \beta_K x_{tK} - \beta_K \rho x_{t-1,K} + u_t. \end{aligned} \quad (3.101)$$

Rovnicu (3.101) nazývame *Durbinova rovnica*. Durbin navrhol použiť konzistentný *OLS* odhad parametru ρ a tento odhad použiť miesto ρ v Durbinovej rovnici. V prípade, že u_t sú normálne, tento odhad je asymptoticky eficientný.

Očividne existuje množstvo *FGLS* odhadov pre daný model. Avšak všetky tieto odhady sú asymptoticky ekvivalentné. Na malých výberoch sa ale výkonnosť jednotlivých odhadov môže líšiť.

3.3.3 Odhad na malom výbere v AR(1)

Asymptotická ekvivalencia predchádzajúcich odhadov prirodzene vedie k otázke, ktorý odhad použiť v prípade malých výberov, kde nemôžeme použiť asymptotické výsledky. Rao a Griliches skúmali *FGLS* odhadov na malom výbere v modele s autokorelovanými chybami pomocou experimentu Monte Carlo. K vygenerovaniu dát pre experiment použili model

$$\begin{aligned} y_t &= x_t + e_t, \\ x_t &= \lambda x_{t-1} + \nu_t, \quad |\lambda| < 1, \\ e_t &= \rho e_{t-1} + u_t, \quad |\rho| < 1, \quad t = 1, 2, \dots, 20. \end{aligned} \quad (3.102)$$

Pre dané ρ a λ , bolo vytvorených 50 vzoriek. Rozptyly ν_t a u_t boli upravené tak, aby $R^2 \doteq 0.9$. Skúmané typy odhadov boli konkrétne: *OLS* odhad, Cochranova–Orcuttova transformácia s Cochranovým–Orcuttovým $\hat{\rho}$ a s Durbinovým $\hat{\rho}$, Praisova–Winstenova transformácia s Cochranovým–Orcuttovým $\hat{\rho}$ a s Durbinovým $\hat{\rho}$ a nelineárne obmedzený *OLS* odhad β a ρ v Durbinovej rovnici. Posledná metóda je asymptoticky ekvivalentná s maximálnou vierohodnosťou.

Hlavné závery tejto práce boli:

- (i) *OLS* odhad je najmenej účinným odhadom spomedzi všetkých porovnávaných metód pre stredne veľké a veľké ρ ($\rho \geq 0.3$).

- (ii) Použitím *FGLS* odhadov pri stredne veľkom a veľkom ρ získame pomerne veľa, naopak strata, ktorá môže vzniknúť použitím týchto metód je veľmi malá.
- (iii) Všetky *FGLS* odhady majú relatívne podobnú výkonnosť

3.3.4 ARMA procesy v reziduálnej zložke regresného modelu

Korelácia medzi chybovou zložkou v modele

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kde $\mathbf{E}\mathbf{e} = \mathbf{0}$ a $\mathbf{E}\mathbf{e}\mathbf{e}' = \sigma_e^2\boldsymbol{\Omega}$ môžu mať mnoho foriem. Už popísaný proces AR(1) je špeciálnym prípadom. Všeobecnejšia stacionárna korelačná štruktúra môže byť napríklad

$$\sigma_e^2\boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho_1 & \dots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{T-2} \\ \rho_2 & \rho_1 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho_1 \\ \rho_{T-1} & \rho_{T-2} & \dots & \rho_1 & 1 \end{bmatrix},$$

kde $\rho_k = \mathbf{E}(e_t e_{t-k})/\sigma_e^2$ pre $k = 1, 2, \dots$. Proces, ktorý sme práve popísali je centrováný a korelácie medzi jednotlivými chybami (realizáciami procesu) sú funkciou časového posunu. Jedná sa teda o (slabo) *stacionárny proces*.

Konkrétne stacionárne procesy, ktoré majú pre Boxovu–Jenkinsovu metodológiu výrazný praktický význam sú:

(a) Autoregresný proces rádu p , AR(p)

$$e_t = \phi_1 e_{t-1} + \dots + \phi_p e_{t-p} + u_t, \quad (3.103)$$

kde $\mathbf{E}u_t = 0$, $\mathbf{E}u_t^2 = \sigma_u^2$ a $\mathbf{E}u_t u_{t-k} = 0$, $t = 1, 2, \dots$; $k = 1, 2, \dots$

pomocou operátoru spätného posunutia môžeme model zapísať ako

$$\phi(B)e_t = u_t,$$

kde $\phi(B) = 1 - \sum_{j=1}^p \phi_j B^j$ je *autoregresný operátor*.

Proces AR(p) má rozptyl $\sigma_e^2 = \sigma_u^2/(1 - \phi_1\rho_1 - \dots - \phi_p\rho_p)$.

Autokorelačná funkcia procesu AR(p) je lineárnou kombináciou klesajúcich geometrických postupností a sinusoíd rôznych frekvencií s geometricky klesajúcimi amplitúdami. Parciálna autokorelačná funkcia procesu AR(p) má bod useknutia $k_0 = p$. Proces AR(p) je vždy invertibilný, nakoľko zápis modelu je už v invertovanom tvare.

(b) **Proces kľavých súčtov rádu q , $MA(q)$**

$$e_t = u_t + \theta_1 u_{t-1} + \dots + \theta_q u_{t-q}, \quad (3.104)$$

pomocou operátoru spätného posunutia môžeme model zapísať ako

$$e_t = \theta(B)u_t,$$

kde $\theta(B) = 1 + \sum_{j=1}^q \theta_j B^j$ je *operátor kľavých súčtov*.

Rozptýl procesu $MA(q)$ je

$$\sigma_e^2 = (1 + \theta_1^2 + \dots + \theta_q^2) \sigma_u^2$$

a autokorelačná funkcia je nenulová iba pre prvých q posunov v čase

$$\rho_k = \frac{\theta_k + \sum_{t=1}^q \theta_t \theta_{t+1}}{1 + \sum_{t=1}^q \theta_t^2}, \quad k = 1, \dots, q.$$

Proces $MA(q)$ má konečnú nenulovú korelačnú štruktúru.

(c) **Zmiešaný proces rádu p a q , $ARMA(p,q)$**

$$e_t = \phi_1 e_{t-1} + \dots + \phi_p e_{t-p} + u_t + \theta_1 u_{t-1} + \dots + \theta_q u_{t-q}. \quad (3.105)$$

Pomocou operátoru spätného posunutia môžeme model zapísať ako

$$\phi(B)e_t = \theta(B)u_t.$$

Proces $ARMA(p,q)$ je zmiešaným procesom $AR(p)$ a $MA(q)$, takže jeho vlastnosti, a teda aj vlastnosti autokorelačnej funkcie prejavujú vlastnosti oboch procesov. Autokorelačná funkcia procesu $ARMA(p,q)$ nemá bod useknutia a je lineárnou kombináciou klesajúcich geometrických postupností a sinusoíd rôznych frekvencií s geometricky klesajúcimi amplitúdami, ale s výnimkou prvých $q - p$ hodnôt. Parciálna autokorelačná funkcia procesu $ARMA(p,q)$ taktiež nemá bod useknutia a je ohraničená lineárnou kombináciou klesajúcich geometrických postupností a sinusoíd rôznych frekvencií s geometricky klesajúcimi amplitúdami s výnimkou prvých $p - q$ počiatočných hodnôt.

Analogicky ako u $AR(1)$ *OLS*-odhad parametrov v $ARMA(p,q)$ procesoch nedáva moc dobré výsledky. Preto sa v $ARMA(p,q)$ procesoch obvykle používa buď nelineárne obmedzený *OLS*-odhad alebo maximálne vierochný odhad. Oba odhady sú asymptoticky eficientné za danej normality chýb.

Ako už bolo spomenuté, $ARMA$ procesy sú teoretickým východiskom, pre konštrukciu Boxových–Jenkinsových modelov. Konštrukcia týchto modelov je však opísaná v množstve zahraničnej i českej a slovenskej literatúry. Preto iba v stručnosti popíšeme dôležité kroky.

Konštrukcia Boxových–Jenkinsových modelov

Ako už bolo spomenuté, konštrukcia modelov podľa metodológie, ktorú zaviedli Box a Jenkins sa obvykle vykonáva v troch krokoch

I Identifikácia modelu: Jedná sa o odhadnutie konkrétneho typu procesu na základe vlastností skúmaných dát, napríklad výberová autokorelačná funkcia nemá bod useknutia a parciálna autokorelačná funkcia má bod useknutia $k_0 = 2$, takže pre časový rad identifikujeme model AR(2).

II Odhad modelu: odhadneme konkrétne parametre identifikovaného modelu.

III Diagnostika modelu: Odhadnutý model testujeme na hladine významnosti 5 %, alebo posúdime kvalitu predpovedí (predikcií) modelu.

V prípade, že pri diagnostike testu zistíme, že model nedáva uspokojivé výsledky, musíme všetky tri kroky zopakovať a odhadnúť nový model, toto opakujeme, až kým nenájde uspokojivý model. Väčšinou sa ale pri hľadaní nového modelu jedná len o korekciu parametrov modelu, ktorý sme zamietli.

Identifikácia modelu

V identifikácii modelu zohráva obrovskú rolu tvar autokorelačnej a parciálnej autokorelačnej funkcie. V prvom kroku spočítame autokorelácie a parciálne autokorelácie skúmaného stacionárneho radu a zobrazíme ich graficky (graf korelácií sa obvykle nazýva korelogram). Prezrieme si oba korelogramy a odhadneme bod useknutia pre ρ_k i ρ_{kk} . Je nutné poznamenať, že výberové autokorelácie zrejme nebudú za bodom useknutia rovné nule, ale budú „dostatočne blízko nule.“ Pričom pod pojmom „dostatočne blízko nule“ myslíme, že na základe výberových autokorelácií r_k a r_{kk} môžeme predpokladať, že ρ_k a ρ_{kk} sú za bodom useknutia nulové. K tomu nám poslúži *Bartlettova*, resp. *Quenouilleova aproximácia*. To znamená, že ak pre všetky $k > k_0$ platí, že r_k nepresahuje hodnotu Bartletovej aproximácie, resp. r_{kk} nepresahuje hodnotu Quenouilleovej aproximácie, pričom pre k_0 ešte tieto hodnoty presahovali, môžeme bod k_0 prehlásiť za bod useknutia. Väčšina štatistických software do korelogramu štandardne túto aproximáciu zakresľuje. Takže určenie bodu useknutia je pomerne jednoduché a je založené na vzhľade korelogramu. Ďalším prístupom pre identifikáciu modelu je identifikácia na základe informačných kritérií. Tento prístup umožňuje automatickú identifikáciu. Princípom identifikácie na základe informačného kritéria je nájsť model, pre ktorý je informačné kritérium minimálne. Hodnoty informačných kritérií obvykle závisia na počte parametrov modelu. Dve najpoužívanéjšie informačné kritériá sú *Akaikeho informačné kritérium*, označované ako *AIC* a *Bayesovo informačné kritérium* - *BIC*. Kritériá AIC a BIC pre výber modelu ARMA(p,q) majú tvar:

$$AIC(k, l) = \ln \hat{\sigma}_{k,l}^2 + \frac{2(k + l + 1)}{n} \quad (3.106)$$

a

$$BIC(k, l) = \ln \hat{\sigma}_{k,l}^2 + \frac{(k + l + 1) \ln n}{n}, \quad (3.107)$$

kde $\hat{\sigma}_{k,l}^2$ je odhadnutý rozptyl bieleho šumu procesu ARMA(k,l).

Odhad modelu modelu

Ako už bolo spomenuté, v ARMA(p,q) modeloch sa obvykle používa buď nelineárne obmedzený *OLS*-odhad alebo maximálne vierohodný odhad. Nelineárne najmenšie štvorce a maximálne vierohodné odhady dávajú veľmi blízke výsledky. Pri samotnom modelovaní však nemusíme nutne poznať algoritmy pre výpočet týchto odhadov, nakoľko snád' všetky súčasné štatistické software majú implementované procedúry pre odhadovanie stacionárnych modelov Boxovej–Jenkinsovej metodológie, a teda odhady sú čiste softwarovou záležitosťou. Softwarový výpočet nám zvyčajne poskytne okrem samotných odhadov parametrov aj príslušné smerodajné odchýlky odhadov.

Overovanie modelu

Overovanie modelov Boxovej–Jenkinsovej metodológie je pomerne prepracované a prebieha vo viacerých krokoch. Výsledkom overovania je, buď potvrdená alebo zamietnutá adekvátnosť odhadnutého modelu. V prípade zamietnutia modelu, sa musíme vrátiť späť k identifikácii modelu. Odhadnutý model môžeme overovať z rôznych hľadísk a s rôznou účinnosťou, preto sa doporučuje používať viac metód overovania modelu súčasne. Jednotlivé aspekty ktoré je dobré overovať sú:

- *Súhrnné korelačné chovanie sa odhadnutých reziduí modelu.* Jedná sa o nejakú formu tzv. *Portmanteau testu*, nazývaného tiež *Q-test*, ktorý testuje nulovú hypotézu, že zvolený model je adekvátny. Test prebieha pomocou súhrnného korelačného chovania prvých K odhadnutých reziduí a vďaka jednoduchosti výpočtu je veľmi často používaný. Za K sa obvykle doporučuje voliť približne \sqrt{n} . Najčastejšie sa u Portmanteau testov používa buď *Boxova–Pierceova štatistika* alebo *Ljungova–Boxova štatistika*. Boxova–Pierceova štatistika má tvar

$$Q = n \sum_{k=1}^K (r_k(\hat{\epsilon}_t))^2 \quad (3.108)$$

a Ljungova–Boxova štatistika má tvar

$$Q = n(n+2) \sum_{k=1}^K \frac{(r_k(\hat{\epsilon}_t))^2}{n-k}, \quad (3.109)$$

pričom pre obe štatistiky platí, že pri väčšom rozsahu výberu majú približne χ^2_{K-p-q} rozdelenie.

- *Individuálne chovanie sa odhadnutých reziduí modelu.* Odhadnuté reziduá (resp. odhadnutý biely šum) sú pre diagnostiku modelu veľmi dôležité. Zvyčajne pre ne očakávame, normalitu, nekorelovanosť, homoskedasticitu a nulovú strednú hodnotu, takže diagnostika bude založená na overovaní týchto vlastností. Väčšinu týchto vlastností dokážeme odhadnúť na základe grafických výstupov

štatistického software. Takže kontrolujeme Q–Q graf, histogram, korelogram a pod. Normalitu však môžeme overovať aj na základe nejakého bežne používaného formálneho testu normality, napr. Kolmogorovov–Smirnovov test alebo Shapirov–Wilkov test, vid’ napr. [12].

- *Kontrola ARMA štruktúry.* Korelačná štruktúra odhadnutá z dát by mala približne zodpovedať korelačnej štruktúre odhadnutého modelu.
- *Kontrola parametrov odhadnutého modelu.* V prípade, že smerodajné odchýlky odhadnutých parametrov sú neúmerne veľké, alebo ak odhadnutý rozptyl chýb má veľkú hodnotu, mali by sme skúsiť model preparametrizovať. Doporučuje sa použiť väčšie množstvo parametrov. Ak budú smerodajné odchýlky nového modelu výrazne menšie a odhady nových parametrov budú výrazne nenulové, nahradíme pôvodný model novým modelom.
- *Kontrola stacionarity.* Odhadnutý model by mal spĺňať podmienky stacionarity.

3.3.5 Procesy vykazujúce nekonštantný trend, zmenu úrovne alebo zmenu trendu

Reálne časové rady väčšinou nie sú stacionárne. Často vykazujú nejaký trend, ktorý sa navyše môže v čase meniť, prípadne sa v určitých bodoch zmení úroveň časového radu. Popis takýchto procesov často umožňujú modely typu ARIMA, ktoré stochasticky modelujú aj trend. Pre konštrukciu ARIMA teda nepotrebujeme, aby modelovaný časový rad bol stacionárny, ale je vyžadované aby rad bolo možné stacionarizovať diferencovaním. Pozrime sa teda na definíciu spomínaného procesu ARIMA:

Proces ARIMA(p,d,q)

Integrovaný zmiešaný proces rádu p, d, q ARIMA(p,d,q) definujeme ako

$$\phi(B)w_t = \alpha + \theta(B)\epsilon_t, \quad (3.110)$$

kde

$$w_t = \Delta^d y_t, \quad (3.111)$$

je d –tá diferencia modelovaného časového radu $\{y_t\}$. A ďalej platí, že (3.110) je stacionárny model ARMA(p,q) pre proces w_t .

Pretože

$$\Delta y_t = y_t - y_{t-1} = (1 - B)y_t, \quad (3.112)$$

diferenčný operátor Δ je možné vyjadriť pomocou operátoru spätného posunutia B ako

$$\Delta = 1 - B. \quad (3.113)$$

ARIMA(p,d,q) môžeme teda zapísať aj v súhrnnom tvare ako

$$\phi(B)\Delta^d y_t = \alpha + \theta(B)\epsilon_t. \quad (3.114)$$

Výstavba modelu ARIMA je založená na určení rádu diferencovania a na následnej konštrukcii ARMA na stacionarizovaný rad. Pred diferencovaním ale musíme zistiť či časový rad nie je nutné linearizovať nejakou transformáciou, aby sme ho mohli modelovať pomocou procesu ARIMA, ktorý patrí do skupiny lineárnych modelov časových radov. Box a Cox navrhli veľmi všeobecnú transformáciu, ktorá má tvar:

$$\begin{aligned} y_t^{(\lambda)} &= \frac{y_t^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ y_t^{(\lambda)} &= \log y_t, & \lambda = 0, \end{aligned} \quad (3.115)$$

kde $y_t^{(\lambda)}$ je transformovaný rad. Neznáma hodnota λ sa hľadá pomocou komplikovanej optimalizačnej procedúry založenej na princípe maximálnej vierohodnosti. Táto konštrukcia je výpočtovo náročná. Praktickejší prístup uviedol Jenkins v príručke [9], keď navrhol mocninovú transformáciu

$$\begin{aligned} y_t^{(\lambda)} &= y_t^\lambda, & \lambda \neq 0, \\ y_t^{(\lambda)} &= \log y_t, & \lambda = 0. \end{aligned} \quad (3.116)$$

O hodnote λ rozhodneme graficky. Sledovaný rad rozdelíme na krátke úseky o dĺžke 4 až 12 pozorovaní a v každom úseku spočítame aritmetický priemer m a rozdiel maximálnej a minimálnej hodnoty úseku r . Takto spočítané dvojice zanesieme do grafu, pričom na x -ovej ose budú hodnoty m a na y -ovej ose budú hodnoty r . Pokiaľ takto skonštruované body ležia približne na vodorovnej priamke, nezávisí lokálna variabilita radu na úrovni radu, takže nemusíme transformovať. Odhad λ touto metódou je veľmi hrubý, ale to nám nevadí, niektorí autori dokonca doporučujú rozlišovať len medzi hodnotami 0 a 1 pre λ . Určenie λ môžeme obmedziť na štyri prípady. Pre prípad, že graf vyzerá ako konštantná funkcia, položíme $\lambda = 1$. Pre rastúci lineárny trend grafu bude $\lambda = 0$, v prípade konvexnej rastúcej funkcie $\lambda = -0,5$ a v prípade konkávnej rastúcej funkcie $\lambda = 0,5$.

Po identifikovaní potreby transformovať pôvodný časový rad a po prípadnej transformácii musíme určiť rád diferencovania. V praxi sa málokedy používa rád väčší než 2. V prípade, že pri pohľade na graf časového radu máme pochybnosti o stacionarite, vykreslíme si korelogram výberovej autokovariančnej funkcie radu a v prípade, že hodnoty autokorelačnej funkcie klesajú pomaly, t.j. približne lineárne, je to príznakom, že musíme diferencovať. Tento postup opakujeme až kým nedostaneme stacionárny rad. Pre stacionarizovaný rad následne vybudujeme model ARMA(p, q).

3.3.6 Procesy so sezónnou zložkou

Podobne ako Boxova–Jenkinsova metodológia zvláda modelovanie trendu, dokáže modelovať aj sezónnosť. Najčastejšie používaným nástrojom k modelovaniu sezónnosti je v Boxovej–Jenkinsovej metodológii *multiplikatívny sezónny proces rádu* $(p, d, q) \times (P, D, Q)_S$, ktorý sa väčšinou nazýva SARIMA rádu $(p, d, q) \times (P, D, Q)_S$ a má tvar

$$\phi(B)\Phi(B^S)\Delta^d\Delta_S^D y_t = \theta(B)\Theta(B^S)\epsilon_t, \quad (3.117)$$

kde ϵ_t je biely šum. Model je vlastne multiplikatívnym spojením dvoch modelov ARIMA. Takže SARIMA obsahuje dva typy autoregresných operátorov (t.j. bežný a sezónny autoregresný operátor) a dva typy operátorov klzavých súčtov (opäť bežný a sezónny). Bežná autoregresná zložka je reprezentovaná autoregresným operátorom (t.j. polynómom $\phi(B)$ rádu p) a bežná zložka klzavých súčtov je reprezentovaná operátorom $\theta(B)$ rádu q . Podobne $\Phi(B^S)$ je sezónny autoregresný operátor rádu P a $\Theta(B^S)$ je sezónny operátor klzavých súčtov rádu Q a nakoniec $\Delta^d = (1 - B)^d$ a $\Delta^D = (1 - B)^D$ označujú bežný a sezónny diferenčný operátor (v tomto poradí). Výstavba sezónnych modelov je o niečo zložitejšia ako výstavba doteraz uvedených modelov aj keď prebieha v troch rovnakých krokoch. Výstavbu modelu SARIMA podrobne popíšeme na príklade v aplikačnej časti.

3.3.7 Predpovede v Boxovej–Jenkinsovej metodológii

Konstruktia predpovedí v Boxovej–Jenkinsovej metodológii je jednoduchá. Aj tento fakt prispieva k obľúbenosti týchto metód. Predpoveď hodnoty y_{t+k} v čase t , t.j. *predpoveď v čase t o k krokov dopredu* označme $\hat{y}_{t+k}(t)$. Princíp výpočtu predpovedí spočíva v minimalizácii strednej štvorcovej chyby $MSE = E(y_{t+k} - \hat{y}_{t+k}(t))^2$. Takže lineárnu predpoveď $\hat{y}_{t+k}(t)$ skonštruujeme ako lineárnu funkciu y_t, y_{t-1}, \dots , ktorá bude mať najmenšiu strednú štvorcovú chybu MSE . V praxi je však výpočet ešte jednoduchší a pre model ARMA(p, q) definovanom pomocou (3.105) môžeme použiť rekurentný výpočet výpočet:

$$\hat{e}_t = \phi_1 \hat{e}_{t-1} + \dots + \phi_p \hat{e}_{t-p} + \hat{\epsilon}_t + \theta_1 \hat{e}_{t-1} + \dots + \theta_q \hat{e}_{t-q}, \quad (3.118)$$

kde $\hat{e}_{t+j}(t) = e_{t+j}$ pre $j \leq 0$, $\hat{e}_{t+j}(t)$ je nulové pre $j > 0$ a pre $j \leq 0$ platí $\epsilon_{t+j} = e_{t+j} = e_{t+j} - \hat{e}_{t+j}(t+j-1) = u_t(t-1)$. Vo výpočte postupujeme vždy o jeden krok dopredu a použijeme odhadnuté parametre modelu. Intervalovú predpoveď pre e_{t+k} v čase t spočítame z bodových predpovedí a z predpokladu normality ako interval s krajnými bodmi $\hat{e}_{t+k}(t) \pm 2$ -násobok odhadnutej smerodajnej odchýlky chýb predpovede, ktorá je rovná rozdielu $e_{t+k} - \hat{e}_{t+k}(t)$.

3.3.8 Autoregresné modely volatility

Nedávne a súčasné problémy vo finančnom sektore naprieč Amerikou a Európou motivujú k štúdiu volatility časových radov v súčasnosti ešte viac ako v minulosti. Prvé modely volatility vznikli už v osemdesiatych rokoch minulého storočia. Prielomovým modelom bol model *autoregresnej podmienenej volatility* ARCH, ktorý použil Engle už v roku 1982 na modelovanie inflácie vo Veľkej Británii. Jedná sa vlastne o priamu aplikáciu Boxovej–Jenkinsovej metodológie pre volatilitu. Model ARCH rádu m , t.j. ARCH(m) ktorý vyjadruje volatilitu ako jednoduchú kvadratickú funkciu omeškaných hodnôt reziduí môžeme zapísať v tvare:

$$y_t = \mu_t + e_t, \quad e_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \sum_{i=0}^m \alpha_i e_{t-i}^2, \quad (3.119)$$

kde ϵ_t sú *i.i.d.* s nulovou strednou hodnotou a jednotkovým rozptylom. Podmienená stredná hodnota μ_t sa často modeluje pomocou regresného modelu alebo ARMA procesu. Koeficienty α_i musia spĺňať isté podmienky regularity tak, aby e_t mali konečné nepodmienené momenty aspoň do druhého rádu. Ako postačujúca podmienka sa často požaduje $\alpha_0 > 0$ a $\alpha_i \geq 0$ pre $i = 1, 2, \dots, m$.

Rád modelu ARCH určíme jednoducho ako bod useknutia parciálneho korelogramu štvorcov chybovej zložky e_t . V prípade, že sa nepreukážu korelácie medzi štvorcami chýb, resp. formálny test (napr. LM-test pre testovanie ARCH efektov, v R `ArchTest()`) nepreukáže prítomnosť ARCH efektov, ARCH model nie je vhodný.

Zrejme najúspešnejším nástrojom na modelovanie volatility je *zobecnený model autoregresnej podmienenej volatility* GARCH a jeho modifikácie. Model GARCH odstraňuje niektoré nedostatky modelu ARCH ako napríklad vysoký rád m , z čoho vyplýva veľký počet parametrov alebo nezohľadnenie asymetrií. Model GARCH(m, s) má tvar

$$y_t = \mu_t + e_t, \quad e_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \sum_{i=0}^m \alpha_i e_{t-i}^2 + \sum_{j=1}^s \beta_j \sigma_{t-j}^2, \quad (3.120)$$

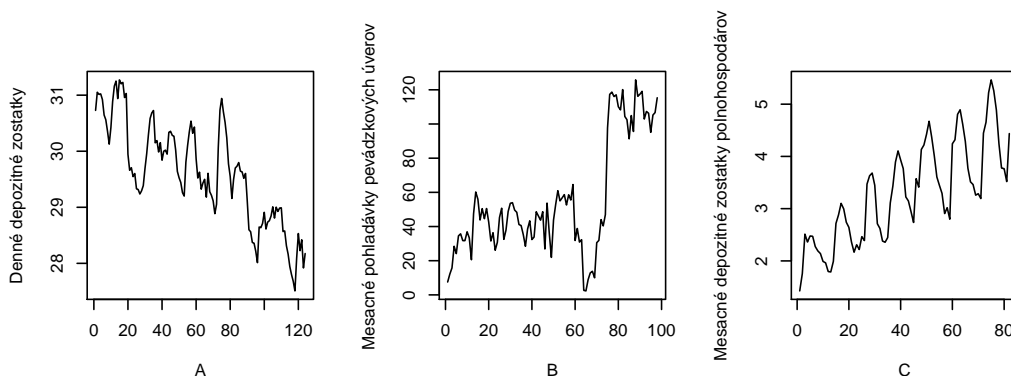
kde ϵ_t sú *i.i.d.* s nulovou strednou hodnotou a jednotkovým rozptylom a parametre modelu spĺňajú $\alpha_0 > 0, \alpha_i \geq 0, \beta_j \geq 0, \sum_{i=1}^{\max\{m, s\}} (\alpha_i + \beta_j) < 1$. Rovnica volatility GARCH je vlastne ARMA pre rad štvorcových odchýliek e_t^2 .

Kapitola 4

Aplikácia na reálnych dátach

4.1 Popis dát

Teoretické poznatky sú demonštrované na príkladoch, v ktorých používam tri rôzne časové rady. Všetky časové rady sú z finančného prostredia a jedná sa buď o denné alebo o mesačné časové rady z bankového prostredia. Vybrala som si však buď náhodné výbery klientov alebo určité segmenty portfólia, a to z dôvodu ochrany interných informácií poskytovateľa dát. Všetky tri časové rady, s ktorými sa budeme ďalej zaoberať sú zobrazené na obrázku 4.1 a sú nahrané na CD s diplomovou prácou. Podkladové dáta sú taktiež depersonifikované. Denné dáta nie sú equidis-



Obr. 4.1: Skúmané časové rady.

tantné, nakoľko k zaúčtovaniu zmien dochádza iba počas pracovných dní. Navyše zostatky z víkendových a sviatočných dní sú vždy zhodné so zostatkami v najbližšom predchádzajúcom pracovnom dni. Umelé pridanie voľných dní by do časových radov pridalo krátke konštantné, resp. nulové úseky, ktoré sú pre portfóliové chovanie sa neprirozené. Z tohto dôvodu sú voľné dni pre modelovanie a predikovanie nezaujímavé a nežiadúce, preto neequidistantné intervaly u denných dát nevadia.

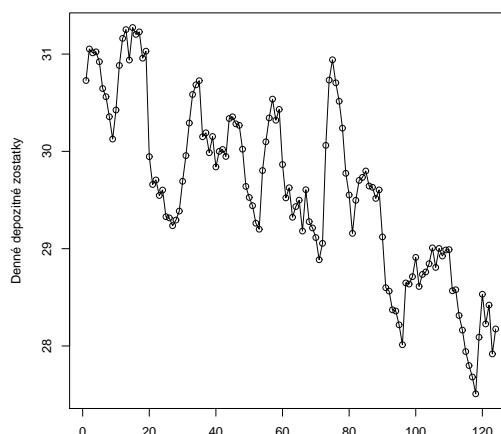
Prvý sledovaný časový rad (obrázok 4.1A) je rad denných zostatkov na vkladoch náhodného výberu firemných klientov banky naprieč rôznymi sektormi oboru podnikania i veľkosti firmy. Sledované obdobie časového radu je 124 pracovných dní, čo je

v skutočnosti obdobie cca 6 mesiacov. Druhý časový rad (obrázok 4.1B) je mesačný a znázorňuje stav pohľadávok (NEA – *net earning assets*) prevádzkového financovania u náhodne vybraných malých klientov za obdobie 104 mesiacov a posledný časový rad (obrázok 4.1C) zobrazuje opäť depozitné zostatky, tentokrát však v mesačnej frekvencii a na náhodnom výbere z poľnohospodárskych klientov. V poslednom časovom rade je sledovaným obdobím 82 mesiacov. Všetky uvedené rady sú buď v miliónoch alebo v miliardách českých korún, ale nakoľko jednotky časového radu nie sú pre samotné modelovanie dôležité, opäť ich z dôvodu ochrany dát neuvádzame. Príklady sú spracované pomocou voľne dostupného štatistického software R.

4.2 Spracovanie a analýza dát

4.2.1 Príklad 1

Zaoberajme sa najprv časovým radom znázorneným na obrázku 4.1A, resp detailnejšie na obrázku 4.2.



Obr. 4.2: Časový rad denných depozitných zostatkov

Ako je vidieť z grafického záznamu, rozhodne sa nejedná o stacionárnu časovú radu, nakoľko jej priebeh sa v čase mení. V takomto prípade by sme mali rozlíšiť, či sa jedná o systematické alebo nsystematické zmeny v úrovni časového radu. Nsystematické zmeny by nás viedli k integrovaným ARMA procesom (t.j. k ARIMA) a museli by sme pristúpiť k diferencovaniu časového radu, na ktorý by sme následne aplikovali ARMA model. Avšak v prípade, že je zmena systematická a v časovej rade vieme rozlíšiť deterministický trend v literatúre sa môžeme dočítať (napr. v [5]), že by sme sa mali pokúsiť trend z časového radu eliminovať. V [5] je uvedené, že keď v tomto prípade prejdeme k diferencovaniu časového radu miesto, eliminácie trendu, dostaneme časový rad s reziduálnou zložkou v tvare MA procesu, aj napriek tejto výstrahe však je možné pomocou Boxovej–Jenkinsovej metodológie modelovať aj

systematický trend. Keďže v tomto prípade sa zmena úrovne radu javí ako deterministický trend, ktorý by sa dal popísať nejakou vhodnou klesajúcou funkciou, mali by sme pristúpiť zrejme k eliminácii trendu. Schválne však preskúmame obe možnosti pre stacionarizovanie časového radu.

Z grafu časového radu je jasné, že eliminovať nebudeme lineárny, ale skôr nejaký mocninový klesajúci trend. V prístupe pomocou eliminácie trendu sa nám ponúka vyskúšať polymomiálny trend (keďže lineárny sme na základe vykreslenia dát zamietli). Pri skúmaní rady sme vyskúšali použiť polynómy 2. a 3. rádu, konkrétne najprv som vyskúšala plný model druhého rádu, ktorý má tvar $y_t = \beta_0 + \beta_1 t + \beta_2 t^2$ a označíme ho ako V1. Pomocou R sa ale koeficient pri lineárnom člene neukázal ako signifikantný, v tomto prípade však nie sú splnené predpoklady klasickej regresie, takže na štatistickú významnosť jednotlivých koeficientov nepozeralme tak striktné. Ide nám iba o elimináciu trendu. Skúsme však radšej lineárny člen vynechať a na dáta aplikujeme ešte variantu polynómu druhého stupňa bez lineárneho člena v tvare $y_t = \beta_0 + \beta_1 t^2$, ktorý označíme V2. Tieto dva lineárne modely môžeme porovnať pomocou výstupu z R:

V1

```
> summary(f1)
-----
Call:
lm(formula = ts_depo[, 4] ~ t + t2)

Residuals:
    Min       1Q   Median       3Q      Max
-1.10200 -0.38355  0.00635  0.40426  1.48387

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.064e+01  1.440e-01 212.768  <2e-16 ***
t           -8.395e-03  5.314e-03  -1.580  0.1167
t2          -9.745e-05  4.115e-05  -2.368  0.0195 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5259 on 121 degrees of freedom
Multiple R-squared:  0.6734,    Adjusted R-squared:  0.668
F-statistic: 124.7 on 2 and 121 DF,  p-value: < 2.2e-16
-----
```

V2

```
> summary(f2)
-----
Call:
lm(formula = ts_depo[, 4] ~ t2)

Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.08497 -0.41415 -0.00688 0.40740 1.40645
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.044e+01  7.145e-02  425.99  <2e-16 ***
t2           -1.607e-04  1.029e-05  -15.62  <2e-16 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5291 on 122 degrees of freedom

Multiple R-squared: 0.6666, Adjusted R-squared: 0.6639

F-statistic: 244 on 1 and 122 DF, p-value: < 2.2e-16

Vyskúšajme na elimináciu trendu použiť ešte polynóm tretieho stupňa. Tento model bude mať tvar $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$ a túto variantu označíme ako V3. Pomocou R dostaneme opäť výstup:

V3

```
> summary(f3)
```

Call:

```
lm(formula = ts_depo[, 4] ~ t + t2 + t3)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-1.06525 -0.38650  0.03492  0.34759  1.38406
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.101e+01  1.890e-01  164.090  < 2e-16 ***
t            -4.333e-02  1.303e-02   -3.325  0.00117 **
t2           5.978e-04  2.415e-04    2.476  0.01470 *
t3          -3.705e-06  1.269e-06   -2.919  0.00419 **
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5103 on 120 degrees of freedom

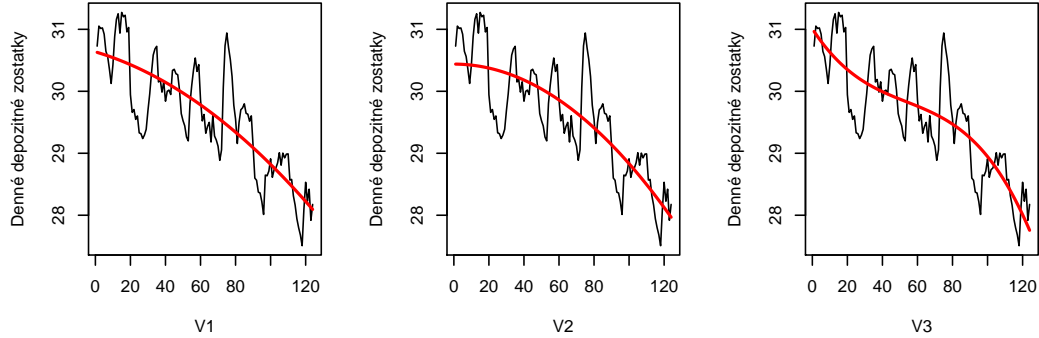
Multiple R-squared: 0.695, Adjusted R-squared: 0.6874

F-statistic: 91.16 on 3 and 120 DF, p-value: < 2.2e-16

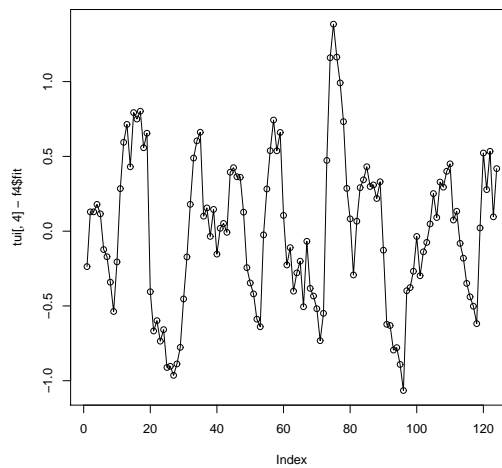
Porovnaním koeficientov determinácie ako aj z grafického zobrazenia trendov do časového radu (viď obrázok 4.3) sa zdá, že varianta V3 popisuje systematický trend v časovom rade najlepšie. Skúsme teda z dát eliminovať trend determinovaný polynómom 3. stupňa.

Elimináciou trendu dostaneme nový časový rad, ktorý už môžeme považovať za stacionárny časový rad. Tento rad je zobrazený na obrázku 4.4 a zobrazme hneď aj jeho korelogram.

Z korelogramu na obrázku 4.5 je zrejmé, že neexistuje (dostatočne malé) k_0 také, že pre všetky $k > k_0$ by sme mohli na základe Bartletovej alebo Quenouilleovej aproximácie prehlásiť autokorelačnú funkciu ρ_k , resp. parciálnu autokorelačnú funkciu



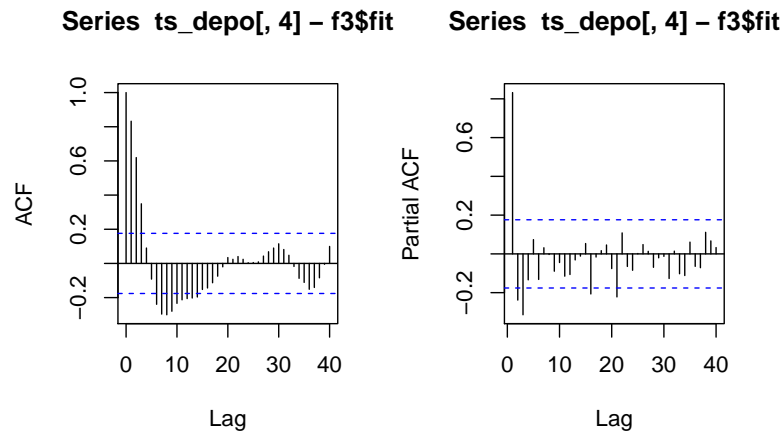
Obr. 4.3: Trendy premietnuté do časového radu depozitných zostatkov



Obr. 4.4: Nový časový rad obdržaný elimináciou trendu

ρ_{kk} za nulový, t.j. nedokážeme nájsť dostatočne malý identifikačný bod k_0 pre autokorelačnú ani pre parciálnu autokorelačnú funkciu procesu. Preto adeptom na proces, ktorý najlepšie popisuje dáta s eliminovaným trendom je ARMA. Skúsme preto spočítať Akaikeho informačné kritérium, označované skrátene tiež AIC kritérium pre rôzne parametre p a q ARMA modelu.

Spočítané AIC kritériá uvedené v tabuľke 4.1 spolu s autokorelačnou a parciálnou autokorelačnou funkciou nasvedčujú tomu, že vhodným modelom by mohla byť ARMA(2,1) alebo ARMA(4,1). Takže pre rad s eliminovaným trendom zostrojíme modely ARMA(2,1) a ARMA(4,1) a skúsime sa ešte pozrieť na konkrétne modely a ich diagnostiku. Aj napriek tvaru autokorelačnej funkcie, ktorá nenasvedčuje modelu AR(p), pozrime sa aj na model AR(4), pretože AIC pre AR(4) vychádza tiež pomerne malé. Parametre modelu AR(4) dostaneme ako výstup z R zodpovedajúci



Obr. 4.5: Autokorelačná (ACF) a parciálna autokorelačná (PACF) funkcia nového časového radu s eliminovaným polynomiálnym trendom

p, q	0	1	2	3	4	5
0		94.11989	59.66365	32.39275	31.24045	26.35284
1	39.74387	36.58664	28.5262	26.85093	28.77721	27.08709
2	33.07545	22.23109	24.45326	27.76135	29.61719	
3	23.79321	24.19896	26.13584	27.53157		
4	23.56545	22.88657	24.36349			
5	24.57895	24.74072				

Tabuľka 4.1: Spočítané hodnoty Akaikeho informačného kritéria pre ARMA(p,q) modely časového radu depozitných zostatkov eliminovaného o systematický trend

modelu ARMA(4,0) aplikovaného na časový rade depozitných zostatkov s eliminovaným trendom:

```
> arma40
```

```
Call:
```

```
arma(x = ts_depo[, 4] - f3$fit, order = c(4, 0, 0))
```

```
Coefficients:
```

```
      ar1      ar2      ar3      ar4 intercept
      0.9276 0.0654 -0.1633 -0.1356    -0.0049
s.e.    0.0896 0.1246 0.1258 0.0904     0.0739
```

```
sigma^2 estimated as 0.06342: log likelihood = -5.78, aic = 23.57
```

Pretože R pracuje s notáciou pre AR(p) v tvare

$$y_t = a_1 y_{t-1} + \dots + a_p y_{t-p} + e_t,$$

model AR(4) má tvar:

$$y_t = 0.9276 y_{t-1} + 0.0654 y_{t-2} - 0.1633 y_{t-3} - 0.1356 y_{t-4} + e_t - 0.0049.$$

Chybovú zložku v práve skonštruovanom AR(4) môžeme považovať za normálnu, pretože na základe Shapirovej–Wilkovej štatistiky ani na základe Kolmogorovej–Smirnovovej štatistiky nemôžeme normalitu zamietnuť. Oba testy sú implementované v R, preto na testovanie normality reziduí sme použili oba testy.

```
> shapiro.test(ar4$residuals)
```

```
-----
```

Shapiro-Wilk normality test

```
data:  ar4$residuals
W = 0.9859, p-value = 0.2298
```

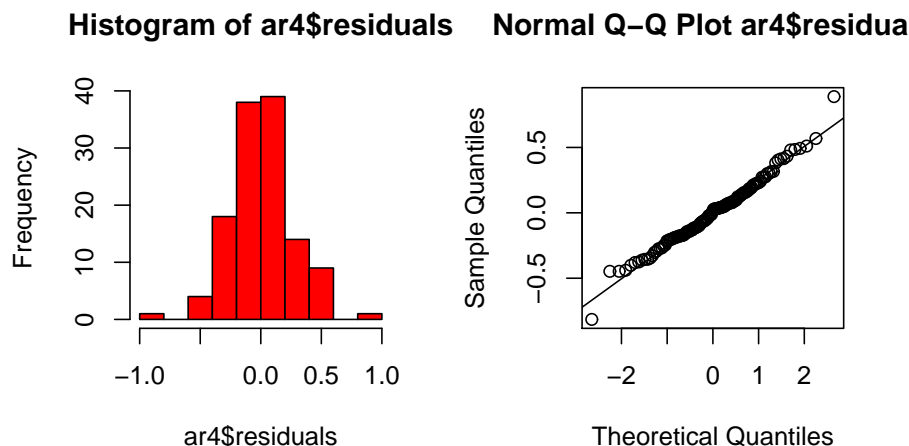
```
> ks.test(ar4$residuals,"pnorm",0,sd(ar4$residuals))
```

```
-----
```

One-sample Kolmogorov-Smirnov test

```
data:  ar4$residuals
D = 0.0466, p-value = 0.9504
alternative hypothesis: two-sided
```

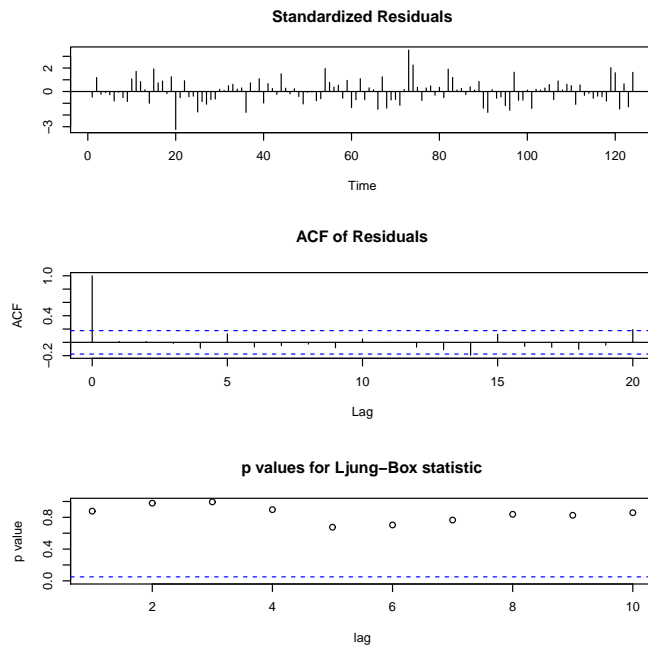
Tomuto faktu zodpovedajú aj histogram a Q–Q graf zobrazený na obrázku 4.6. Napriek tomu, že odhadnutá autokorelačná a parciálna autokorelačná funkcia svojim



Obr. 4.6: Histogram a Q–Q graf reziduí modelu AR(4)

tvarom nasvedčujú skôr modelu ARMA(p,q), AR(4) môžeme použiť. Diagnostiku modelu AR(4) si môžeme prezrieť na obrázku 4.7.

Prejdime pre porovnanie aj k ARMA modelom, ktoré sme podľa tvaru výberovej autokorelačnej a parciálnej autokorelačnej funkcie (a ARMA(4,1) tiež podľa najnižšej hodnoty AIC) považovali za kandidátov na výsledný model. U modelu ARMA(2,1) vychádzajú všetky parametre s výnimkou interceptu signifikantne nenulové, naopak pri pohľade na smerodajné odchýlky koeficientov a samotné koeficienty ARMA(4,1) vidíme, že okrem interceptu, do 95% konfidenčného intervalu spadne 0 aj v prípade koeficientov pri y_{t-1} a y_{t-3} .



Obr. 4.7: Diagnostika modelu AR(4)

```
> arma21
```

```
-----
```

```
Call:
```

```
arima(x = ts_depo[, 4] - f3$fit, order = c(2, 0, 1))
```

```
Coefficients:
```

	ar1	ar2	ma1	intercept
	1.7966	-0.8550	-1.000	0.0010
s.e.	0.0433	0.0432	0.028	0.0102

```
sigma^2 estimated as 0.06228: log likelihood = -6.12, aic = 22.23
```

```
-----
```

Je nutné poznamenať, že R pracuje s nasledujúcou notáciou pre ARMA(p,q):

$$y_t = a_1 y_{t-1} + \dots + a_p y_{t-p} + e_t + b_1 e_{t-1} + \dots + b_q e_{t-q}.$$

Takže ako ARMA(2,1) dostávame:

$$y_t = 1.7966 y_{t-1} - 0.855 y_{t-2} + e_t - e_{t-1} + 0.001$$

Pozrime sa ešte na ARMA(4,1):

```
> arma41
```

```
-----
```

```
Call:
```

```
arima(x = ts_depo[, 4] - f3$fit, order = c(4, 0, 1))
```

```
Coefficients:
```

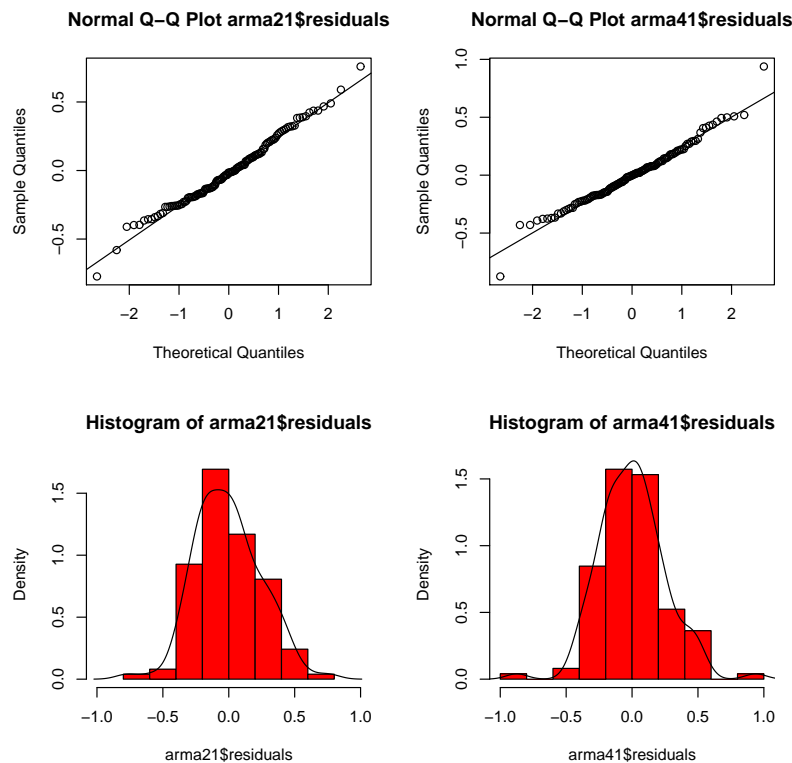
	ar1	ar2	ar3	ar4	ma1	intercept
	0.2789	0.6864	-0.1034	-0.3455	0.6737	-0.0037
s.e.	0.2329	0.2379	0.0945	0.0902	0.2389	0.0772

sigma^2 estimated as 0.06199: log likelihood = -4.44, aic = 22.89

Takže spočítaná ARMA(4,1) má tvar:

$$y_t = 0.2789y_{t-1} + 0.6864y_{t-2} - 0.1034y_{t-3} - 0.3455y_{t-4} + e_t + 0.6737e_{t-1} - 0.0037.$$

Pozrime sa ďalej na reziduá týchto dvoch modelov a na test normality reziduí:



Obr. 4.8: Reziduá ARMA(2,1) a ARMA(4,1)

Na testovanie normality použijeme Shapiro-Wilkov test, ktorý je implementovaný v R:

```
> shapiro.test(arma21$residuals)
```

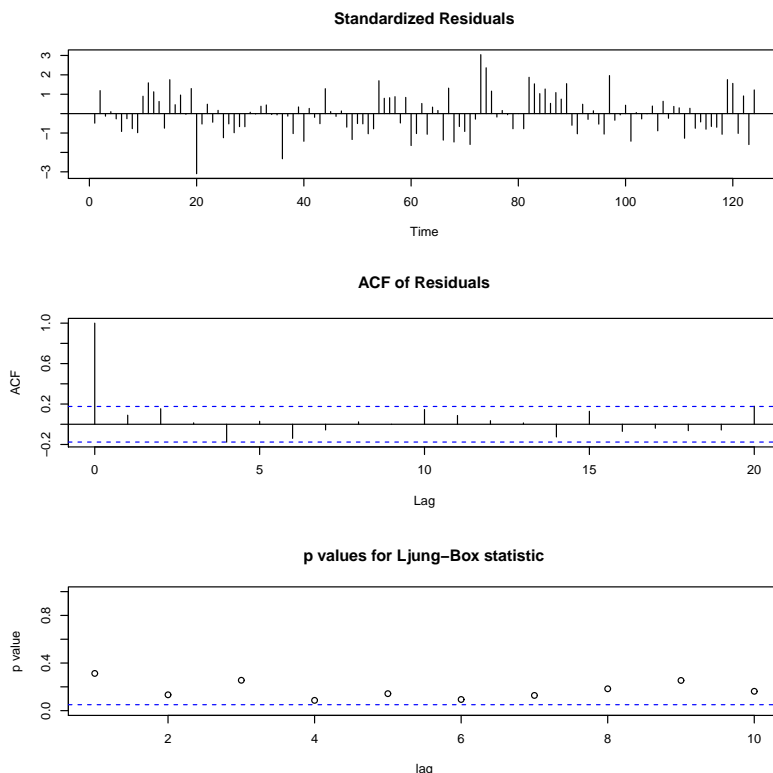
Shapiro-Wilk normality test

```
data: arma21$residuals
W = 0.9884, p-value = 0.3797
```

```
> shapiro.test(arma41$residuals)
```

Shapiro-Wilk normality test

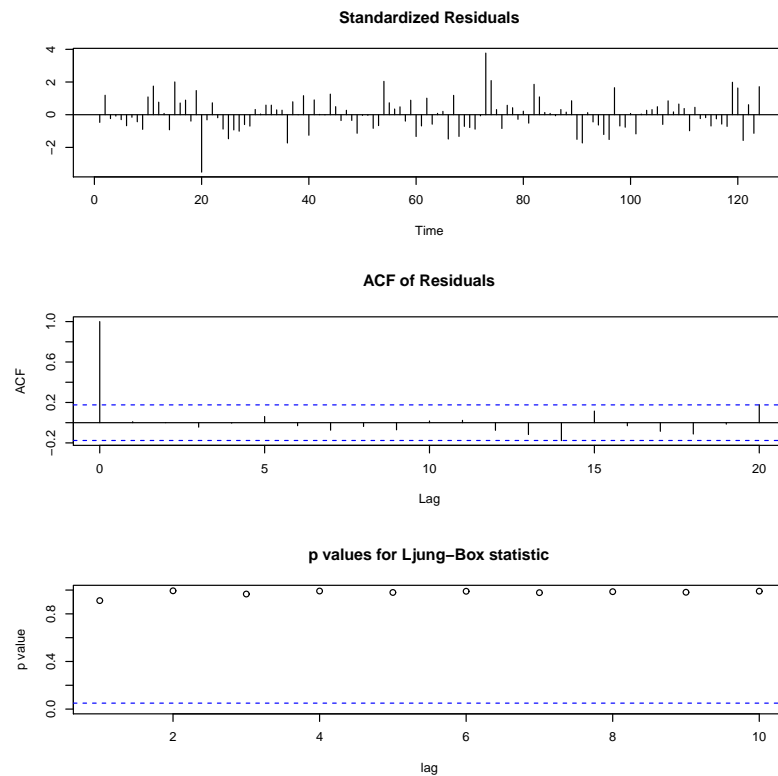
```
data: arma41$residuals
W = 0.9774, p-value = 0.03543
```



Obr. 4.9: Diagnostika ARMA(2,1)

Takže v prípade modelu ARMA(2,1) normalitu na 5% hladine nezamietame, naopak chyby v ARMA(4,1) nemôžeme na 5% hladine považovať za normálne. Reziduá modelu ARMA(2,1) môžeme teda považovať za biely šum. Čo sa týka testovania samotných modelov ARMA, použijeme Ljungovu–Boxovu štatistiku. Z obrázkov 4.9 a 4.10 je vidieť, že na základe Ljungovej–Boxovej štatistiky ani jeden model nezamietame. Vzhľadom k jednoduchosti modelu, signifikantnosti parametrov a reziduám, ktoré tvoria biely šum sa spomedzi modelov ARMA(p,q) jednoznačne prikloníme k modelu ARMA(2,1), pomocou ktorého spočítame aj predikcie. Nakoľko AR(4) spĺňa predpoklad normality zrejme ešte o niečo lepšie ako ARMA(2,1) a p –hodnoty pre Ljungov–Boxov test sú v prípade AR(4) väčšie ako v prípade ARMA(2,1), spočítajme predikcie aj pre AR(4) a pozrime sa ako sa predikcie podľa jednotlivých modelov líšia.

Čo sa týka predikcie, musíme ju vykonať v dvoch krokoch. V prvom kroku vytvoríme odhady 10 nasledujúcich hodnôt časového radu, v ktorom je eliminovaný trend



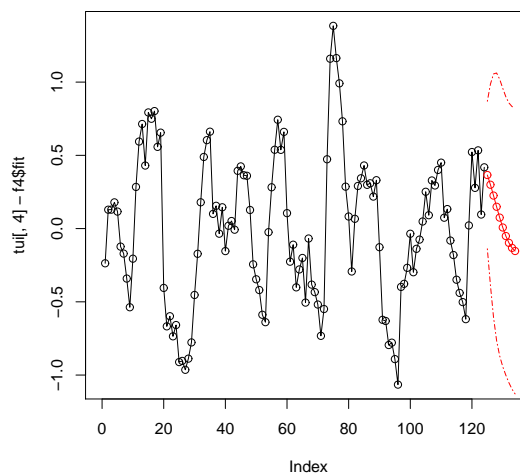
Obr. 4.10: Diagnostika ARMA(4,1)

a v druhom kroku musíme zohľadniť pôvodný trend časového radu. Na obrázkoch 4.11 a 4.12 sú viditeľné malé rozdiely v odhadovaných budúcich hodnotách (obzvlášť v zmene trendu). Avšak pri návrate k pôvodnému časovému radu sa rozdiely v predikciách modelov javia ako minimálne.

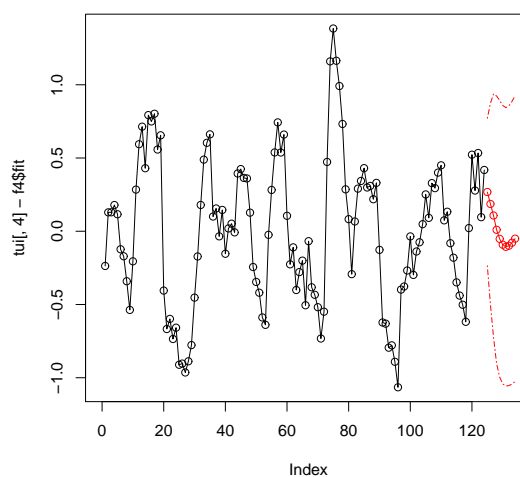
Predikcie k modelu AR(4) vyšli:

```
-----
$pred
Time Series:
Start = 125
End = 134
Frequency = 1
[1] 0.267833917 0.186243359 0.107472249 0.009938702 -0.051975045
[6] -0.091862270 -0.106305913 -0.098980822 -0.078224390 -0.050726002

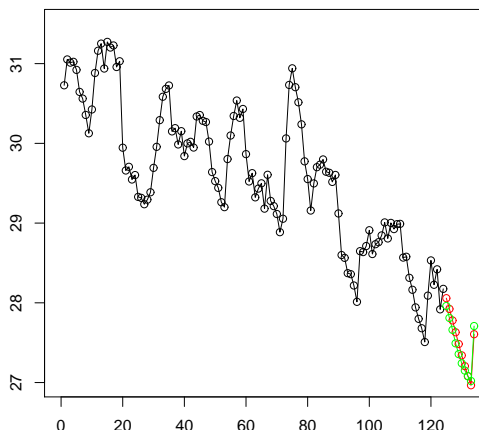
$se
Time Series:
Start = 125
End = 134
Frequency = 1
[1] 0.2518309 0.3434899 0.4151463 0.4567378 0.4721391 0.4751790 0.4752064
[8] 0.4774827 0.4823721 0.4878697
-----
```



Obr. 4.11: Časový rad s eliminovaným trendom s predikciami podľa ARMA(2,1) a 95% konfidenčnými intervalmi



Obr. 4.12: Stacionarizovaný časový depozitných zostatkov s 10 predikciami budúcich hodnôt podľa AR(4)



Obr. 4.13: Predpoveď depozitných zostatkov pre nasledujúcich 10 pracovných dní obdržaná kombináciou trendu a ARMA procesu

Predikcie pre ARMA(2,1):

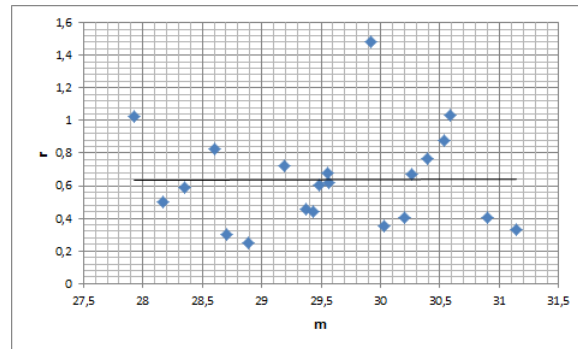
```
-----
$pred
Time Series:
Start = 125
End = 134
Frequency = 1
[1] 0.365661646 0.299430719 0.225377322 0.148959392 0.074981573
[6] 0.007409402 -0.050740561 -0.097439264 -0.131620454 -0.153103337

$se
Time Series:
Start = 125
End = 134
Frequency = 1
[1] 0.2505273 0.3215114 0.3538596 0.3659311 0.3682122 0.3682135 0.3709675
[8] 0.3788034 0.3915398 0.4073539
-----
```

Zodpovedajúce krajné body 95% intervalu spoľahlivosti k tomuto radu spočítame jednoducho vďaka normalite ako predikcia \pm dvojnásobok smerodatnej odchylky predikcie. Takto spočítané predikcie a konfidénčné intervaly sú zobrazené červenou farbou na obrázku 4.11. Nakoniec musíme predpovedať 10 nasledujúcich hodnôt pre model trendu, ktorý sme z časového radu eliminovali. Predpovede trendu spočítame jednoducho pomocou predpisu regresného modelu a výslednú predikciu pre 10 nasledujúcich hodnôt dostaneme ako súčet predikcie polynomiálneho trendu a predikcie modelu ARMA(2,1). Výsledná predpoveď pre oba modely je zobrazená na obrázku 4.13, pre AR(4) červenou a pre ARMA(2,1) zelenou farbou. Nižšia hodnota AIC kritéria, menší počet parametrov ako aj signifikantná nenulovosť všetkých odhad-

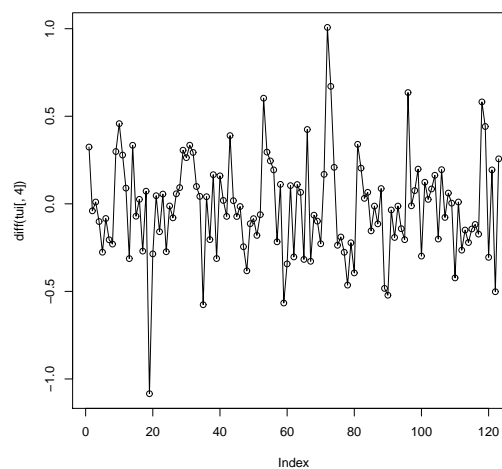
nutých parametrov sú v prospech $\text{ARMA}(2,1)$, aj keď aj distribúcia reziduí je o niečo bližšie k teoretickému normálnemu rozdeleniu u $\text{AR}(4)$. V tomto prípade teda ako výsledný model vyberieme $\text{ARMA}(2,1)$.

△



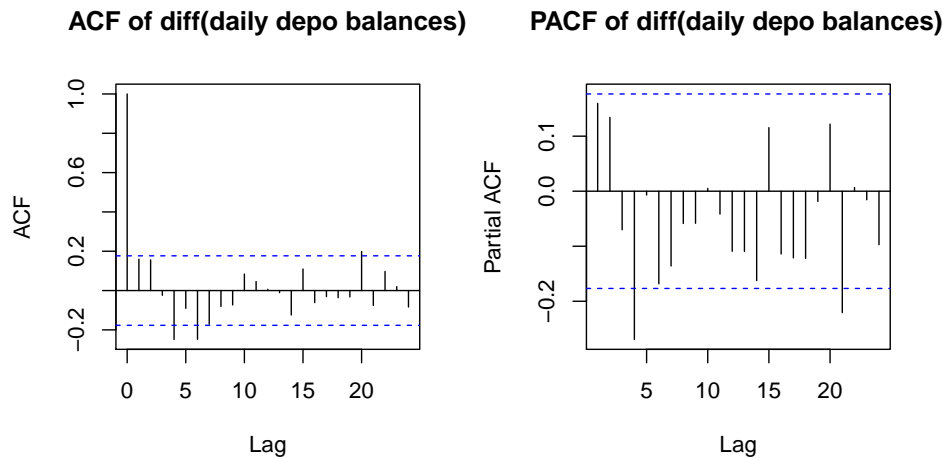
Obr. 4.14: Závislosť úrovne a variability v priebehu časového radu

Prejdime teda k druhému prístupu modelovania časovej rady. V tomto prípade sa pokúsime namodelovať stochasticky aj pozorovaný trend časového radu a kvôli stacionarizácii budeme diferencovať. Pred diferencovaním by sme ale mali overiť, či v časovom rade nedochádza so zmenou úrovne aj ku zmene variability. Pozrime sa teda najprv na graf (obrázok 4.14) aritmetických priemerov a rozdielov maximálnych a minimálnych hodnôt v krátkych úsekoch časového radu, ktoré majú dĺžku 6 pozorovaní. Keďže body na grafe sú približne rovnomerne rozmiestnené okolo konštantného trendu, variabilita časového radu nezávisí na jeho úrovni a rad nebudeme transformovať.



Obr. 4.15: Diferencie 1. rádu

Prejdime k diferenciám časového radu, viď obrázok 4.15. Diferencovaný časový rad má približne nulovú strednú hodnotu, avšak z grafu sa zdá, že v ňom dochádza



Obr. 4.16: Autokorelačná a parciálna autokorelačná funkcia diferencií depozitných zostatkov

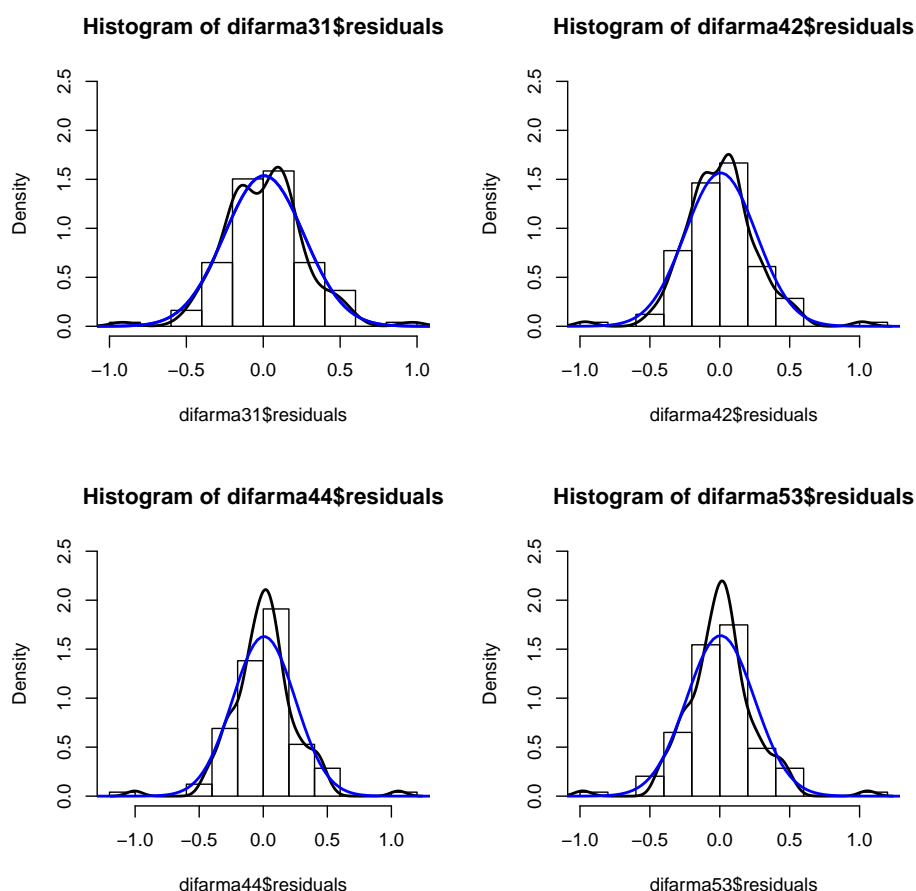
k miernemu zhlukovaniu volatility. Ak by dáta skutočne vykazovali zhlukovanie volatility, nebol by to žiaden problém. Riešením by mohla byť nejaká modifikácia typu ARCH. Skúsme teda nájsť vhodný model pre diferencovaný rad a v prípade, že zistíme prítomnosť ARCH efektu, nájdeme vhodný model typu ARCH/GARCH.

Pretože diferencie depozitných zostatkov sú autokorelované a zrejme nemajú dostatočne malý bod useknutia, viď autokorelačná a parciálna autokorelačná funkcia na obrázku 4.16, hľadáme vhodný ARMA(p,q) model pre rad diferencií. Pozrime sa na Akaikeho informačné kritériá jednotlivých modelov, ktoré sú uvedené v tabuľke 4.2.

p, q	0	1	2	3	4	5
0		46.67052	43.35354	45.24739	41.43614	38.85257
1	45.9413	47.16799	45.31907	44.46914	32.70909	34.50363
2	45.72925	47.59377	39.59123	32.60132	33.35368	35.2783
3	47.04726	31.71658	40.82233	41.98091	34.98451	35.52365
4	39.7426	41.68434	31.4999	31.99588	27.96606	28.81913
5	41.7425	43.36798	32.68046	26.8186	28.79725	30.94943

Tabuľka 4.2: Hodnoty Akaikeho informačného kritéria pre modely ARMA(p,q) diferencovaného časového radu depozitných zostatkov

Z tabuľky 4.2 vidíme, že najnižšie hodnoty Akaikeho informačného kritéria majú modely ARMA(5,3) a ARMA(4,4). Jedná sa však o model s veľkým množstvom parametrov, preto sa pozrieme aj na vlastnosti modelov ARMA(4,2) a ARMA(3,1). Skontrolujeme histogramy (na obrázku 4.17) a Q–Q grafy (na obrázku 4.18) reziduí týchto modelov a tiež prevedieme štandardné testy normality. V histogramoch je čiernou farbou zakreslená empirická hustota a modrou farbou hustota teoretického normálneho rozdelenia so strednou hodnotou a s rozptylom, ktoré sú rovné výberovým stredným hodnotám a výberovým rozptylom reziduí v zodpovedajúcom modeli. Prevedieme formálne testy normality reziduí týchto modelov. Na základe Kolmogorovovho–Smirnovovho testu nemôžeme na 5% hladine zamietnuť normalitu reziduí žiadneho zo skúmaných modelov:



Obr. 4.17: Histogramy reziduí modelov ARMA(p,q) aplikovaných na rad diferencií denných depozitných zostatkov

```
> ks.test(difarma31$residuals,sd(difarma31$residuals) )
```

```
-----  
Two-sample Kolmogorov-Smirnov test
```

```
data: difarma31$residuals and sd(difarma31$residuals)  
D = 0.8699, p-value = 0.2742  
alternative hypothesis: two-sided  
-----
```

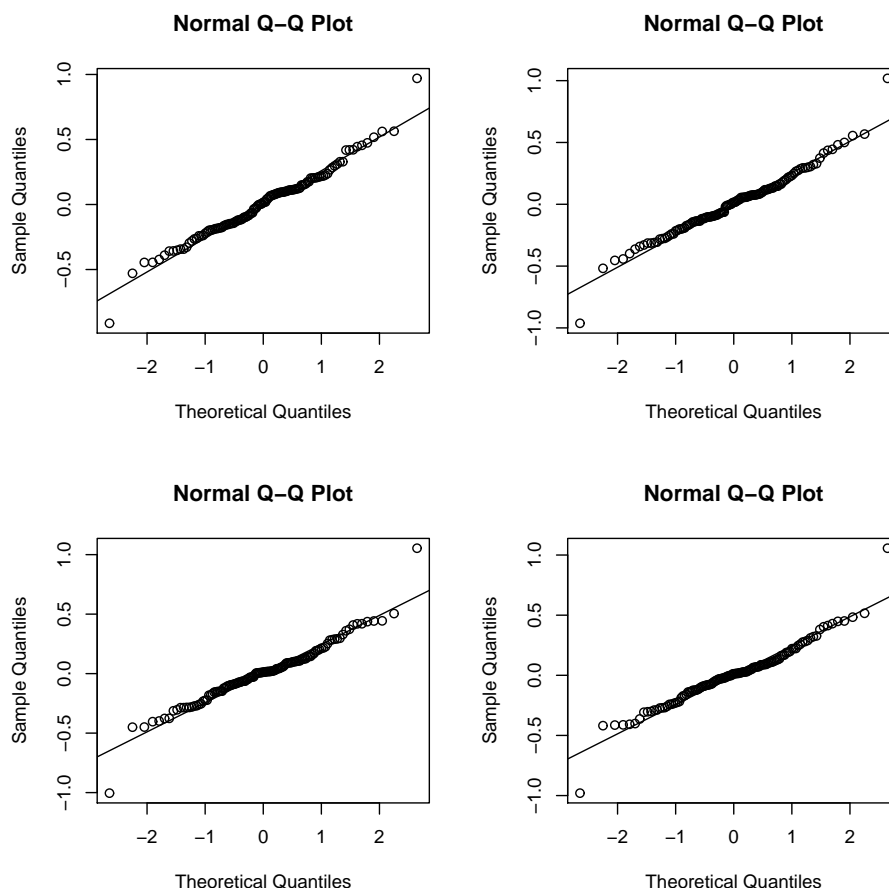
```
> ks.test(difarma42$residuals,sd(difarma42$residuals))
```

```
-----  
Two-sample Kolmogorov-Smirnov test
```

```
data: difarma42$residuals and sd(difarma42$residuals)  
D = 0.8537, p-value = 0.3065  
alternative hypothesis: two-sided  
-----
```

```
> ks.test(difarma44$residuals,sd(difarma44$residuals))
```

```
-----  
Two-sample Kolmogorov-Smirnov test
```



Obr. 4.18: Q–Q grafy reziduí ARMA(p,q) modelov aplikovaných na rad diferencií denných depozitných zostatkov

```
data: difarma44$residuals and sd(difarma44$residuals)
D = 0.8618, p-value = 0.2903
alternative hypothesis: two-sided
```

```
> ks.test(difarma53$residuals,sd(difarma53$residuals))
```

Two-sample Kolmogorov-Smirnov test

```
data: difarma53$residuals and sd(difarma53$residuals)
D = 0.8618, p-value = 0.2903
alternative hypothesis: two-sided
```

Avšak Shapirova–Wilkova štatistika na 5% hladine už tak optimistické výsledky nedáva:

Shapiro-Wilk normality test

```
data: difarma31$residuals
```

W = 0.9781, p-value = 0.04294

```
> shapiro.test(difarma42$residuals)
```

Shapiro-Wilk normality test

data: difarma42\$residuals

W = 0.9692, p-value = 0.006524

```
> shapiro.test(difarma44$residuals)
```

Shapiro-Wilk normality test

data: difarma44\$residuals

W = 0.9508, p-value = 0.0002057

```
> shapiro.test(difarma53$residuals)
```

Shapiro-Wilk normality test

data: difarma53\$residuals

W = 0.9505, p-value = 0.0001935

Na základe Shapirovho–Wilkovho testu zamietame na 5% hladine normalitu reziduí všetkých štyroch sledovaných modelov. Takže prípadné intervalové predpovede odvodené od normálneho rozdelenia budú problematické. Pozrime sa ešte na obrázok 4.19 na autokorlačné funkcie reziduí a štvorcov reziduí v modele ARMA(5,3), ktorý má najnižšiu hodnotu Akaikeho informačného kritéria. Z korelogramov vidíme, že nielen chyby modelu ale aj ich štvorce sú vzájomne nekorelované, čo znamená, že v dátach zrejme nebude žiaden ARCH efekt, čo si overíme ešte formálnym testom v R:

```
> arch.test(difarma53$residuals)
```

ARCH LM-test; Null hypothesis: no ARCH effects

data: difarma53\$residuals

Chi-squared = 2.5831, df = 5, p-value = 0.7639

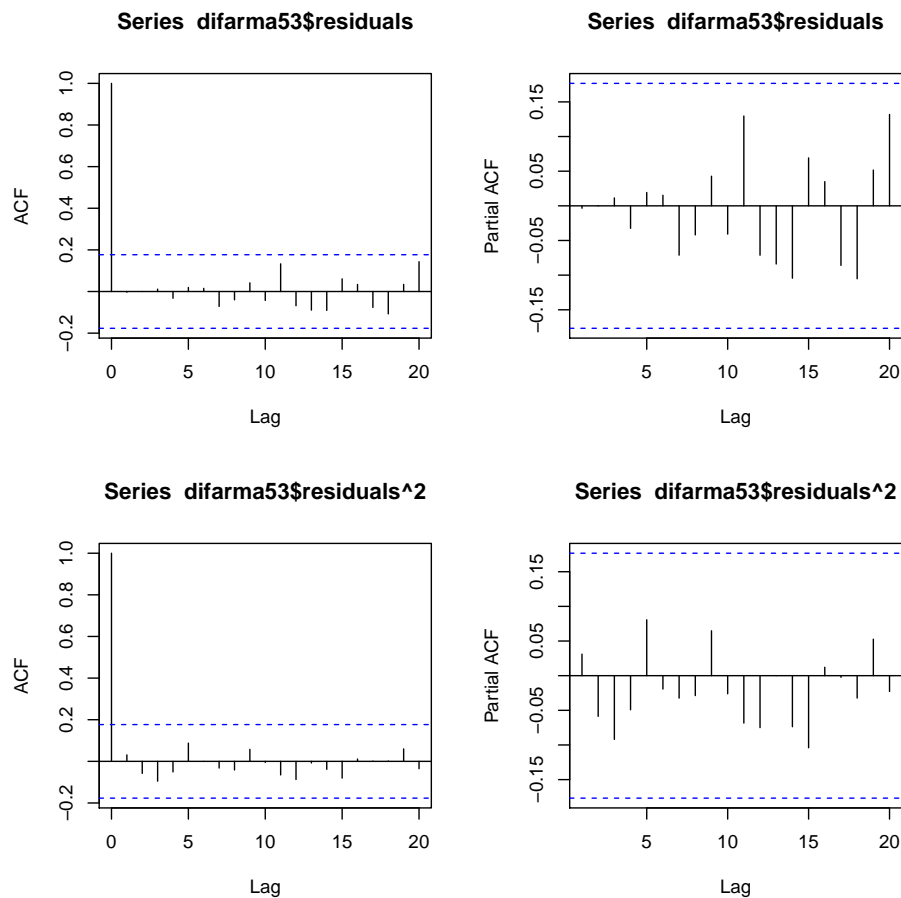
Ako vidíme z výstupu z R, nulovú hypotézu, ktorá tvrdí, že v dátach nie sú žiadne ARCH efekty nemôžeme zamietnuť. Takže naše pôvodné podozrenie na zhukovanú volatilitu sa nepotvrdilo a vystačíme si s ARIMA modelom. Výsledný ARIMA model je teda ARIMA(5,1,3), konkrétne sme z R dostali nasledujúci výstup, z ktorého je jasne vidieť, že všetky parametre modelu sú signifikantne nenulové:

Call:

```
arima(x = tui[, 4], order = c(5, 1, 3))
```

Coefficients:

	ar1	ar2	ar3	ar4	ar5	ma1	ma2	ma3
	-0.5618	0.6395	0.7209	-0.4354	-0.2391	0.7344	-0.4415	-0.8495
s.e.	0.1125	0.1173	0.1036	0.0936	0.0977	0.0837	0.1099	0.0826



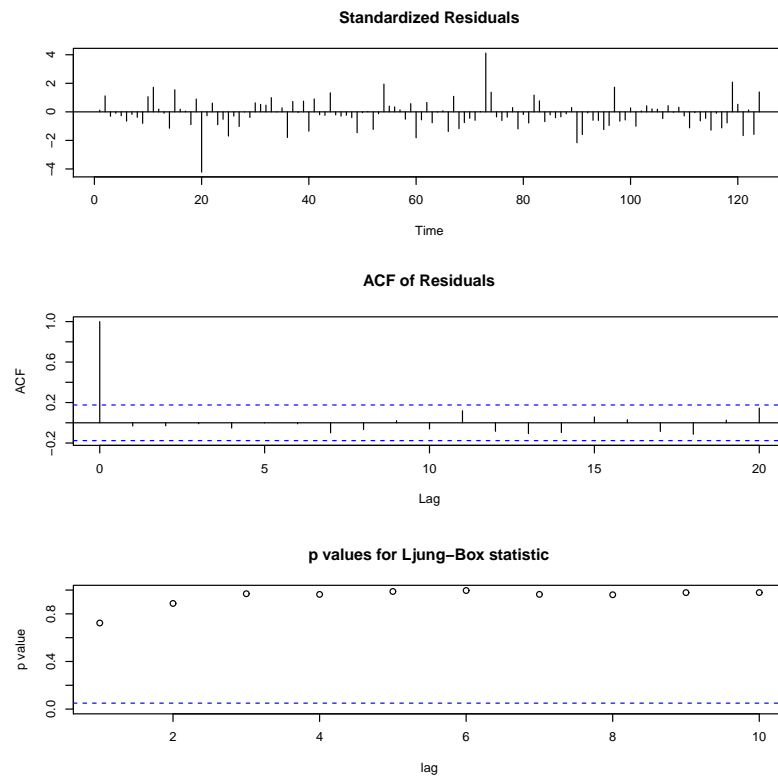
Obr. 4.19: Autokorelačné a parciálne autokorelačné funkcie reziduí a štvorcov reziduí modelu ARMA(5,3)

```
sigma^2 estimated as 0.06413: log likelihood = -7.25, aic = 32.5
```

Skontrolujme ešte diagnostiku modelu ARIMA(5,1,3), vid' obrázok 4.20. Pre porovnanie s prvým prístupom riešenia sa opäť pozrime na predikcie tohto modelu. Konkrétne hodnoty predikcií a ich smerodajné odchýlky spočítané pomocou R sú:

```
$pred
Time Series:
Start = 125
End = 134
Frequency = 1
[1] 28.28475 28.19772 28.37461 28.30690 28.28609 28.39364 28.21488 28.35628
[9] 28.26530 28.23612

$se
Time Series:
Start = 125
```



Obr. 4.20: Diagnostika modelu ARIMA(5,1,3) aplikovaného na denné depozitné zostatky

End = 134

Frequency = 1

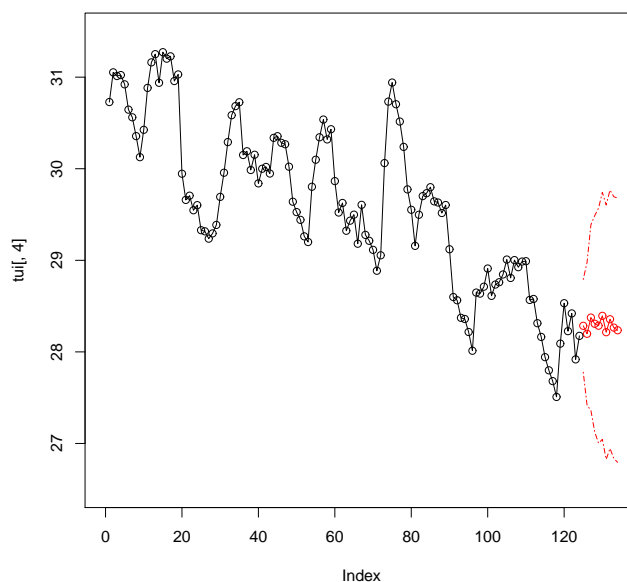
[1] 0.2533762 0.3903239 0.5063553 0.5904063 0.6418256 0.6746014 0.6942535

[8] 0.7054392 0.7146600 0.7210293

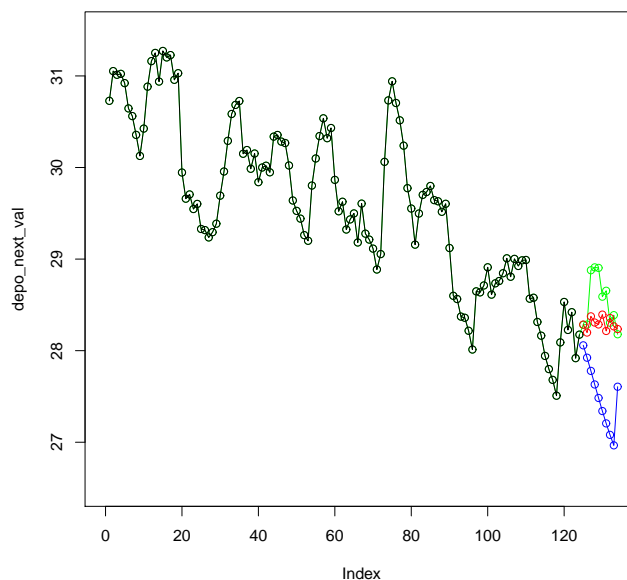
Predikcie spolu s odhadnutými konfidenčnými intervalmi (ktoré, kvôli nenormálne chýb nemusia byť odhadnuté správne) sú zobrazené na obrázku 4.21. Z grafického znázornenia predikcií vidíme, že model, ktorý je vytvorený bez regresnej eliminácie systematického trendu, nepredpokladá, že klesajúci trend bude pretrvávať (aspoň nie pre nasledujúcich 10 hodnôt).

△

Otázkou zostáva ktorý prístup je vhodnejší. Keďže v literatúre sa môžeme dočítať jednak o rizikách spojených s modelovaním systematického trendu pomocou Boxovej–Jenkinsovej metodológie, tak na druhej strane, často sú uvádzané výborné výsledky tejto metodológie aj pri modelovaní trendu alebo sezónnosti. V súčasnosti už máme k dispozícii ďalších 10 skutočných hodnôt časového radu denných depozitných zostatkov. Porovnajme teda oba modely ku ktorým sme dospeli voči realite. Oba typy predikcií ako aj skutočné hodnoty sú zobrazené na obrázku 4.22. Z grafu



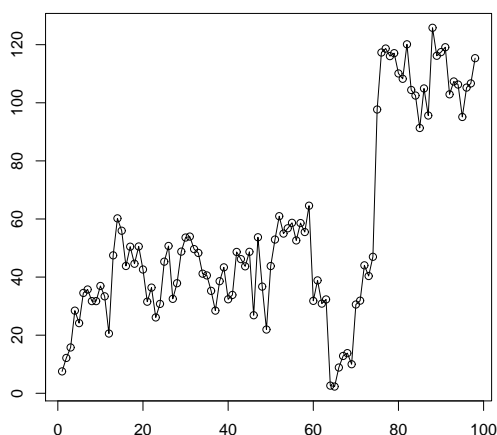
Obr. 4.21: Predikcie modelu ARIMA(5,1,3)



Obr. 4.22: Časový rad denných depozitných zostatkov spolu s predikciami zostrojenými 1.čiste pomocou Boxovej–Jenkinsovej metodológie (červenou farbou), 2. elimináciou trendu a následným použitím Boxovej Jenkinsovej metodológie (modrou farbou) a 3. skutočné budúce hodnoty radu (zelenou farbou)

vidíme, že v realite nielenže klesajúci trend nepokračoval, ale dokonca v priebehu týchto 10 dní zostatky krátkodobo narástli. Jednalo sa o päťdňový výkyv, po ktorom sa reálne zostatky dostali opäť približne na úroveň stochastických predikcií. Čiste stochastický prístup dopadol výrazne lepšie ako prístup v ktorom sme eliminovali vypozerovaný trend. A to aj napriek tomu, že chyby modelu nevykazujú normalitu. Eliminácia trendu by bola zrejme na mieste, keby sme mali konkrétnu expertnú znalosť, ktorá by nasvedčovala tomu, že trend zostane aj naďalej zachovaný. V prípade, že analytik nemá expertnú znalosť modelovanej problematiky a navyše v minulosti časového radu sa dajú vypozerovať (aj keď iba krátkodobé) zmeny trendu, vhodnejší bude zrejme Boxov–Jenkinsov prístup.

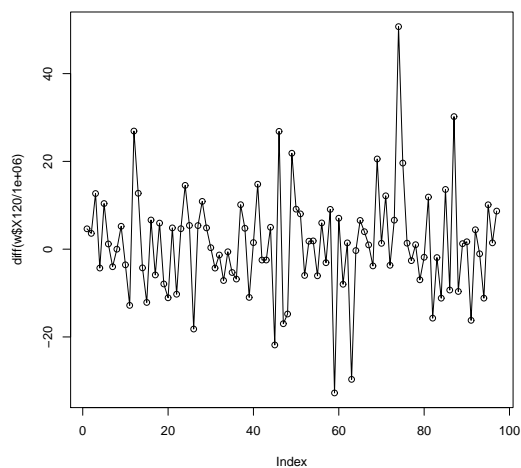
4.2.2 Príklad 2



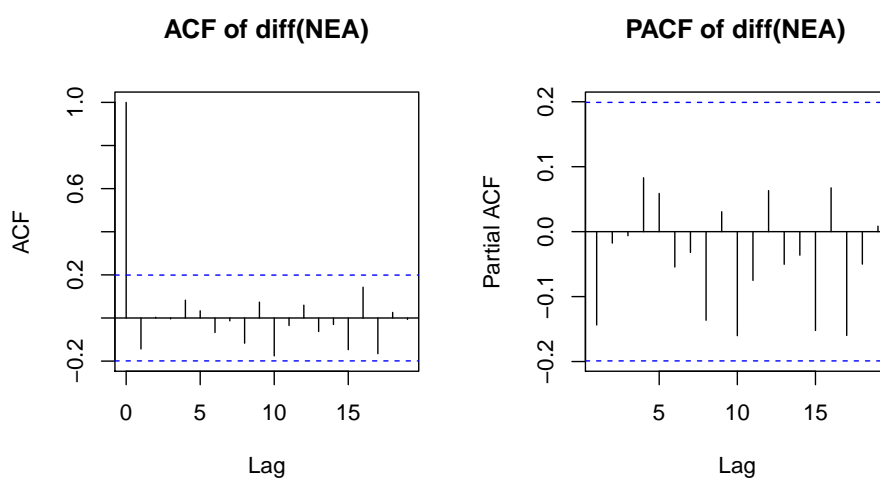
Obr. 4.23: Časový rad pohľadávok na prevádzkovom financovaní

Ďalší skúmaný časový rad je znázornený na obrázku 4.1B, resp. detailnejšie na obrázku 4.23. Opäť sa nejedná o stacionárnu časovú radu, pretože v grafe je jasne rozpoznateľná zmena úrovne a žiaden systematický trend. Stacionarizáciu tohto radu prevedieme diferencovaním. Tento časový rad sa teda pokúsime popísať integrovaným ARMA procesom (t.j. procesom ARIMA(p,d,q)). Výstavbu ARIMA modelu prevedieme v dvoch krokoch. Najprv rad stacionarizujeme diferencovaním a následne na stacionarizovaný rad aplikujeme vhodný ARMA model. Skúsme na rad aplikovať diferencie prvého rádu. Stacionarizáciou obdržíme nový časový rad, ktorý je zobrazený na obrázku 4.24.

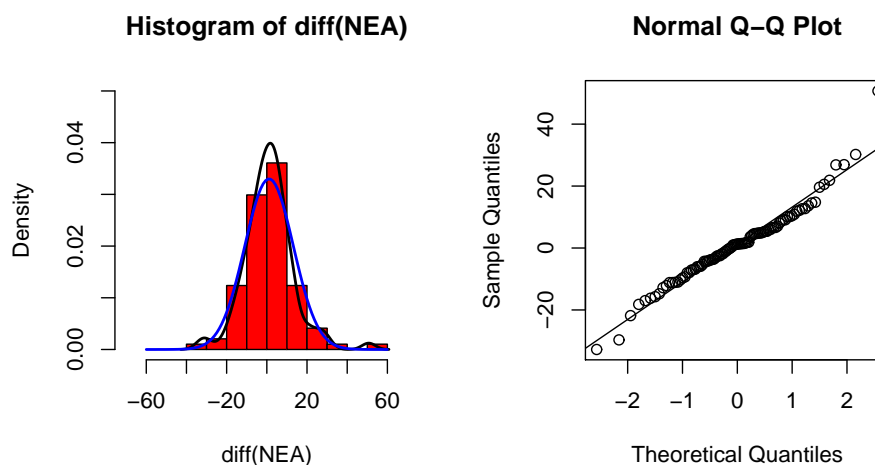
Pozrime sa ešte na korelogram stacionarizovaného radu. Z korelogramu (obrázok 4.25) vidíme, že sa zrejme jedná o náhodnú prechádzku. Overme ešte normalitu diferencovaného radu pomocou Kolmogorovovho–Smirnovovho testu.



Obr. 4.24: Stacionarizovaný časový rad pohľadávok



Obr. 4.25: Autokorelačná a parciálna autokorelačná funkcia diferencovaného radu



Obr. 4.26: Histogram a Q–Q graf diferencovaného časového radu

One-sample Kolmogorov-Smirnov test

```
data: diff(w$X120/1e+06)
D = 0.0926, p-value = 0.3541
alternative hypothesis: two-sided
```

Na základe Kolmogorovovej–Smirnovovej štatistiky normalitu nezamietame, nakoľko p–hodnota je 0.3541. Pretože autokorelačnú a tiež parciálnu autokorelačnú funkciu môžeme na základe Bartletovej, rep. Quenouilleovej aproximácie považovať za nulové (až na rozptyl), tak na pôvodné dáta aplikujeme ARIMA(0,1,0), ktorá má tvar

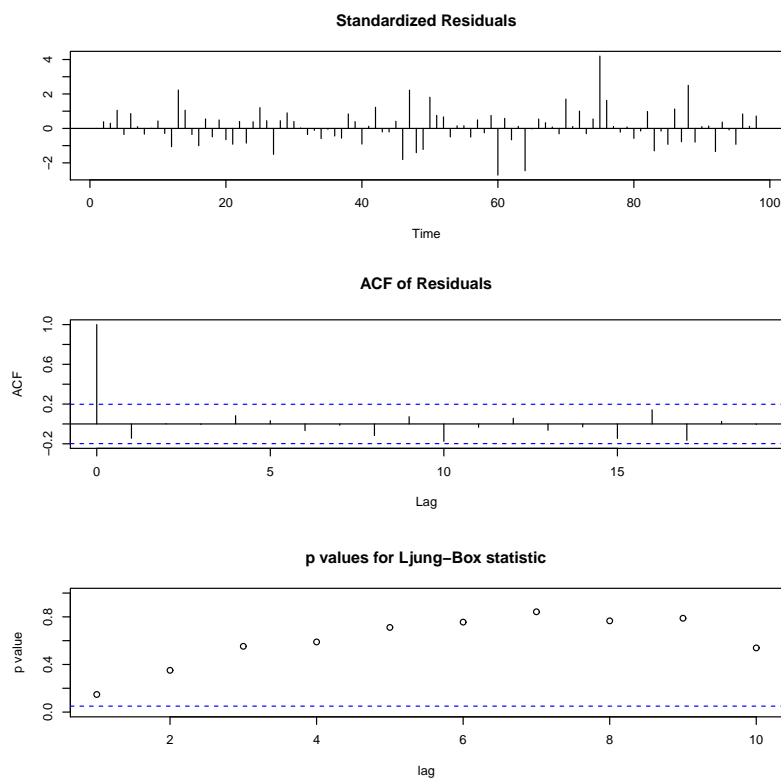
$$y_t - y_{t-1} = e_t.$$

Pri pohľade na diagnostiku modelu zistíme, že reziduá môžeme považovať nielen za normálne rozdelené, ale aj za nekorelované a na základe Ljungovej–Boxovej štatistiky model nemôžeme zamietnuť. Skontrolujme ešte pre istotu Arch efekt pomocou Arch Testu v R:

ARCH LM-test; Null hypothesis: no ARCH effects

```
data: arima010$residuals
Chi-squared = 6.3744, df = 12, p-value = 0.896
```

Nulovú hypotézu, že v dátach sa nenachádzajú žiadne Arch efekty nemôžeme zamietnuť a preto zostaneme pri modeli ARIMA(0,1,0). Nakoniec spočítajme predpovede nasledujúcich hodnôt časového radu a 95% interval spoľahlivosti pre nasledujúce hodnoty. V tomto prípade bude najlepším odhadom pre nasledujúce pozorovanie posledná hodnota časového radu. Takže čo sa týka predikcií do budúcnosti, budú sa líšiť iba v konfidenčných intervaloch, ktoré sa budú smerom do budúcnosti rozširovať, viď nasledujúci výstup z R:

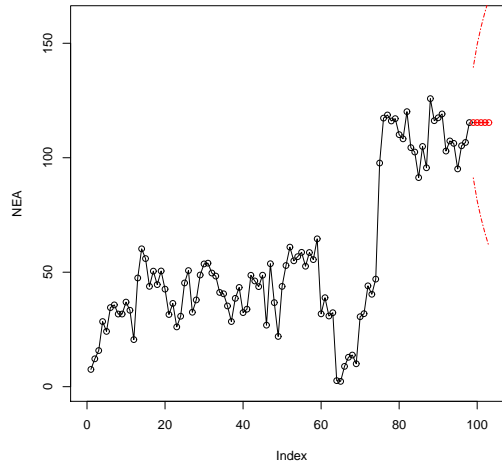


Obr. 4.27: Diagnostika ARIMA(0,1,0)

```
-----
$pred
Time Series:
Start = 99
End = 103
Frequency = 1
[1] 115.3552 115.3552 115.3552 115.3552 115.3552
```

```
$se
Time Series:
Start = 99
End = 103
Frequency = 1
[1] 12.07872 17.08190 20.92096 24.15745 27.00885
-----
```

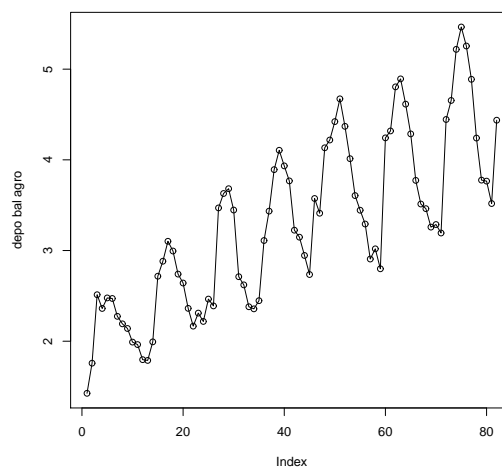
Na obrázku 4.27 je zobrazená diagnostika modelu ARIMA(0,1,0) a konečne, na obrázku 4.28 je zobrazený rad s predikciami a konfidenčnými intervalmi predikcií.



Obr. 4.28: Časový rad pohľadávok s predikciami

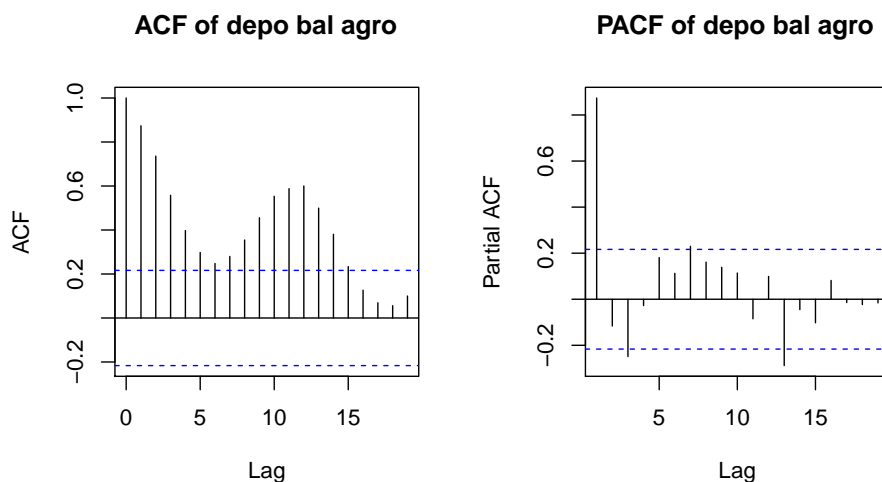
4.2.3 Príklad 3

Posledným skúmaným časovým radom je mesačný rad depozitných zostatkov poľnohospodárov, ktorý je zobrazený na obrázkoch 4.1C a 4.29. V tomto časovom rade je jasne rozpoznateľný rastúci trend a výrazná multiplikatívna sezónnosť (pretože sezónne výkyvy sa s rastúcim trendom zväčšujú). Mohli by sme použiť aj dekom-



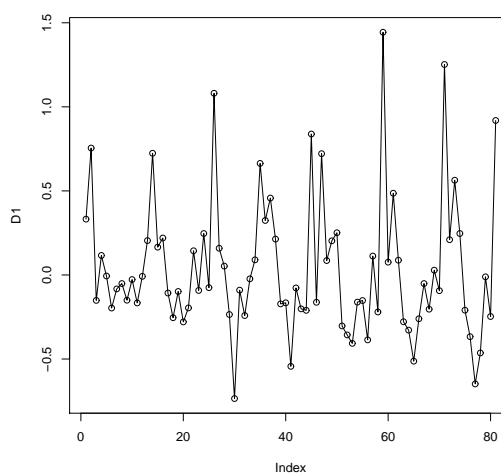
Obr. 4.29: Mesačný časový rad depozitných zostatkov poľnohospodárov

pozičné metódy, avšak chceme sa aj naďalej zaoberať Boxovou–Jenkinsovou metodológiou. Takže sezónnosť i trend skúsime namodelovať stochasticky. Skúsme teda zistiť či dokážeme zostrojiť nejaký vhodný sezónny model Boxovej–Jenkinsovej metodológie, t.j. $SARIMA(p,d,q) \times (P,D,Q)_s$ model. Pri pohľade na graf časového radu zistíme, že dáta vykazujú ročnú sezónnosť, čo zodpovedá aj realite v poľnohospodárstve, nakoľko poľnohospodári zvyknú v našich končinách zasiať alebo investovať väčšinou



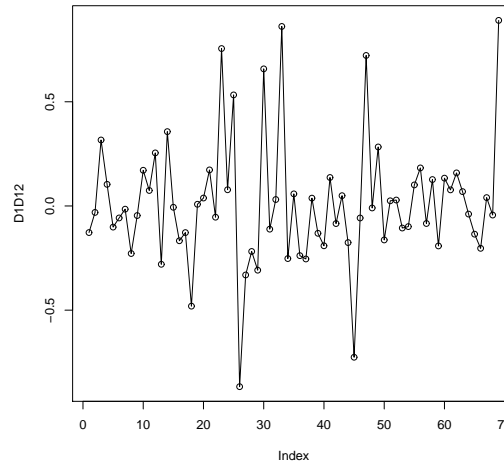
Obr. 4.30: Autokorelačná a parciálna autokorelačná funkcia depozitných zostatkov poľnohospodárov

na prelome zimného a jarného obdobia, keď počasie začne byť vhodné pre pestovanie plodín, do konca jesene pozbierajú a predávajú úrodu, v zime dostanú vyplatené dotácie, v marci väčšinou zaplatia dane za predchádzajúci rok, atď. Vďaka tejto znalosti nemusíme hľadať ďalšie argumenty pre ročnú sezónnosť v dátach, ale keďže cieľom práce je hlavne demonštrácia teoretických znalostí na dátach, pozrieme sa aj na vlastnosti dát, ktoré indikujú konkrétnu sezónnosť. Ročnej sezónnosti zodpovedajú aj korelogramy pôvodného radu. Na obrázku 4.30 vidíme, že autokorelačná i parciálna autokorelačná funkcia nadobúdaajú v bode 12 lokálny extrém, čo je tiež dostatočný argument pre predpoklad ročnej sezónnosti. Takže sa pokúsime nájsť



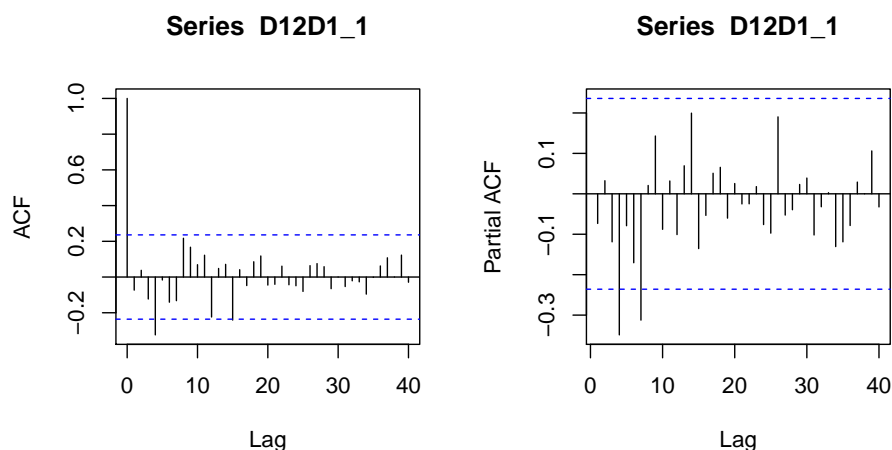
Obr. 4.31: Diferencovaný rad depozitných zostatkov poľnohospodárov

vhodný model s ročnou sezónnosťou, t.j. $s = 12$. Pretože skúmaný časový rad má zjavný rastúci trend, časový rad diferencujeme a tým trend odstránime. Diferenco-



Obr. 4.32: Diferencovaný rad depozitných zostatkov poľnohospodárov po sezónnej diferenciacii

vaný rad je zobrazený na obrázku 4.31. Z radu musíme odstrániť ešte sezónnosť, čo spravíme pomocou sezónnych diferencií. Aplikujme teda na diferencovaný rad sezónne diferencie, viď obrázok 4.32. Sezónne diferencovaný časový rad môžeme považovať za stacionárny a ďalej budeme hľadať $SARIMA(p,d,q) \times (P,D,Q)_s$ model, ktorý čo najlepšie popíše tento stacionarizovaný rad. Keďže vieme, že sezónnosť časového radu je ročná a sezónne diferencovaný rad sme stacionarizovali diferencovaním 1. rádu, môžeme predpokladať, že $d = 1$, $D = 1$ a $s = 12$. Zostáva nám teda určiť p , q , P a Q . Skúmajme teda ďalej autokovariančnú a parciálnu autoko-



Obr. 4.33: Korelačná a parciálna korelačná funkcia stacionarizovaného radu

variančnú funkciu stacionarizovaného časového radu. Pri pohľade na hodnoty autokorelačnej a parciálnej autokorelačnej funkcie zodpovedajúce sezónnemu omeškaniu (t.j. hodnoty autokorelačných funkcií s indexami 12, 24, ...) zistíme, že zodpovedajúce hodnoty ρ_k a ρ_{kk} môžeme považovať za nulové. To znamená, že $P=Q=0$.

Takže nám zostáva určiť p a q . Keďže r_{15} a $r_{6,6}$ sú väčšie ako Bartletova aproximácia, stacionarizovaný proces je zrejme treba odhadnúť pomocou vhodného ARMA procesu, čo znamená, že ani jeden z týchto parametrov nebude nulový. Cieľom je ale odhadnúť pôvodný a nie stacionarizovaný rad. Predpoklad na nenulovosť p a q však zostane zachovaný. Určiť správne parametre modelu SARIMA nie je jednoduchá úloha s exaktným algoritmom riešenia. Nakoľko nám chýbajú k odhadnutiu už iba 2 parametre, vyskúšame radšej preveriť viacero modelov $\text{SARIMA}(p, 1, q) \times (0, 1, 0)_{12}$. Dôležité je, aby reziduá výsledného modelu tvorili biely šum. Preskúmaním viacerých modelov (viď skript na priloženom CD) dostaneme ako vhodný model $\text{SARIMA}(4, 1, 3) \times (0, 1, 0)_{12}$. Konkrétne koeficienty identifikovaného modelu sme dostali ako výstup z R:

```
-----
$fit

Call:
arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q),
period = S), optim.control = list(trace = trc, REPORT = 1, reltol = tol))

Coefficients:
          ar1          ar2          ar3          ar4          ma1          ma2          ma3
      -0.1875  -0.0702  0.3836  -0.3361  -0.0235  -0.0847  -0.8918
s.e.    0.1394   0.1543  0.1227   0.1326   0.0932   0.1571   0.1388

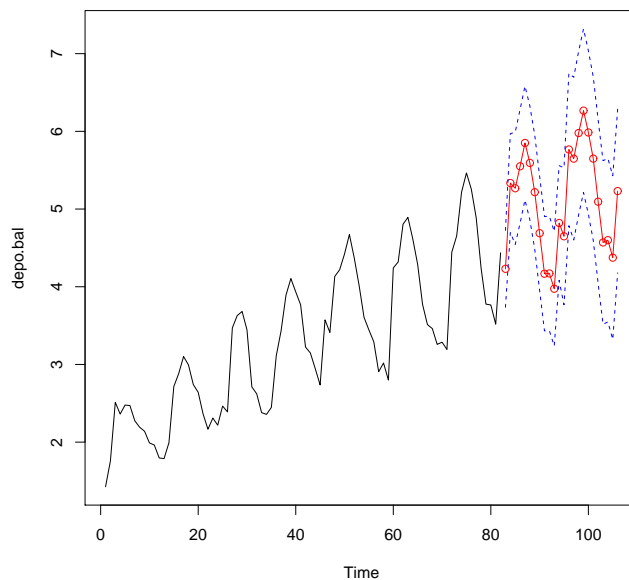
sigma^2 estimated as 0.06043:  log likelihood = -3.59,  aic = 23.18

$AIC
[1] -1.635498

$AICc
[1] -1.587052

$BIC
[1] -2.430047
-----
```

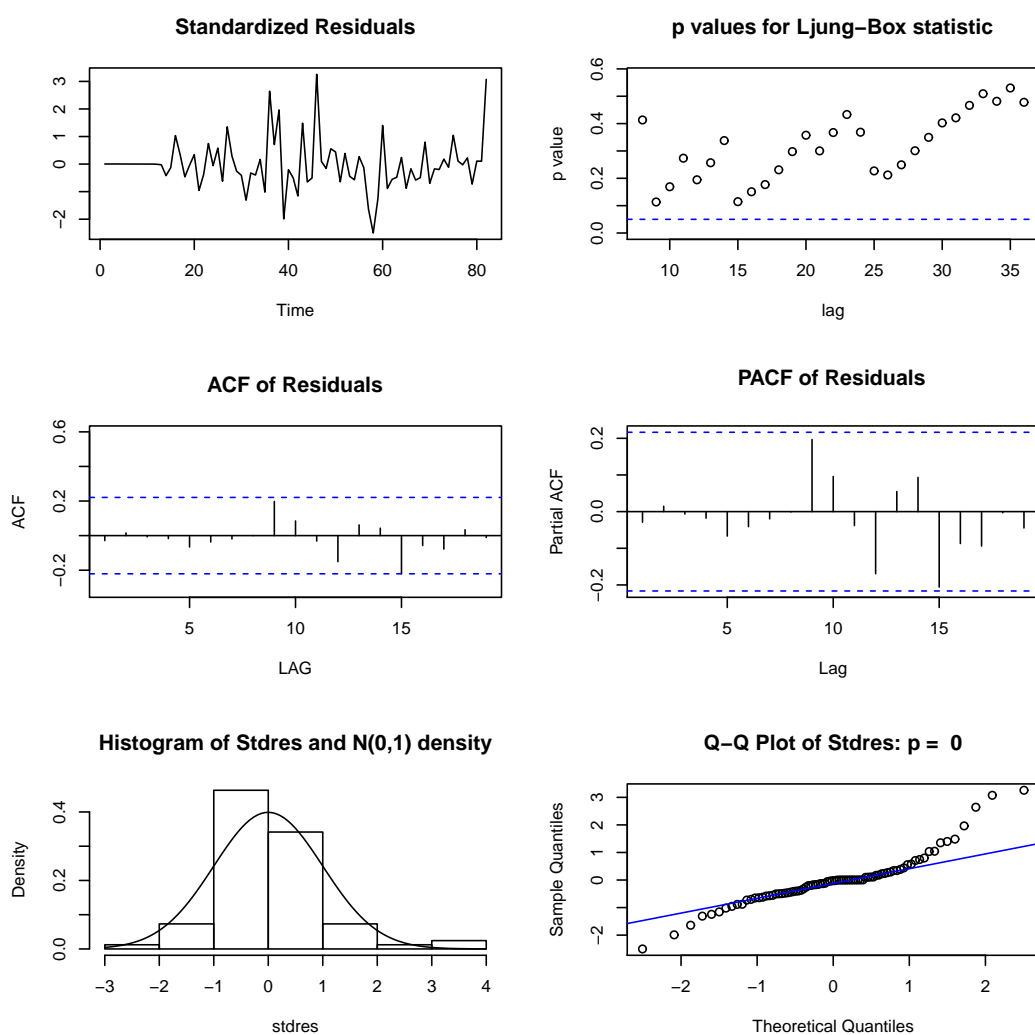
Pri pohľade na výstup z R čitateľa zrejme zarazí záporná hodnota AIC kritéria modelu $\text{SARIMA}(4, 1, 3) \times (0, 1, 0)_{12}$. V dokumentácii k R ako aj v [11] (napr. na str. 153) môžeme nájsť, že napriek s rozporu s teoretickou hodnotou AIC kritéria, ktorá je podľa definície AIC kritéria vždy nenulová, spočítané AIC v R môže byť záporné. Je to z dôvodu odlišného výpočtu informačného kritéria použitého v R, nakoľko R používa pri výpočte odhadnutého AIC kritéria modifikácie teoretickej definície. Diagnostika modelu je zobrazená na obrázku 4.35. Napriek tomu, že v tomto prípade sme dostali iba 4 signifikantne nenulové koeficienty, nebudeme ďalej hľadať vhodnejší model, nakoľko k modelu sme dospeli skúšaním rôznych modelov, spomedzi ktorých model $\text{SARIMA}(4, 1, 3) \times (0, 1, 0)_{12}$ vyšiel najlepšie. Prechod k vyšším rádom by mohol byť už problematický aj z dôvodu softwarových obmedzení, navyše v literatúre sa modely vyšších rádom príliš nedoporučujú. Už teraz sa jedná o pomerne vysoký rád. Nakoniec prejdime k predikciám budúcich hodnôt. Konkrétne predikované hodnoty spolu s ich smerodajnými odchýlkami ako aj grafické znázornenie predikcie (viď obrázok 4.34) dostaneme opäť ako výstup z R. Textový výstup z R:



Obr. 4.34: Pôvodný časový rad s predikciou na nasledujúce dva roky

```
$pred
Time Series:
Start = 83
End = 106
Frequency = 1
 [1] 4.230859 5.334905 5.267214 5.550450 5.850359 5.593800 5.217570 4.689151
 [9] 4.167554 4.171236 3.974386 4.820528 4.649361 5.766262 5.648854 5.978485
[17] 6.266290 5.985351 5.649051 5.094619 4.569812 4.599434 4.374546 5.231636

$se
Time Series:
Start = 83
End = 106
Frequency = 1
 [1] 0.2475551 0.3174373 0.3614225 0.3660141 0.3664585 0.3664718 0.3672905
 [8] 0.3674400 0.3681679 0.3681362 0.3682865 0.3688378 0.4408168 0.4905494
[15] 0.5237176 0.5263732 0.5261010 0.5260276 0.5262680 0.5260318 0.5264693
[22] 0.5262123 0.5266617 0.5269413
-----
```



Obr. 4.35: Diagnostika modelu SARIMA(4, 1, 3) × (0, 1, 0)₁₂ aplikovaného na depozitné zostatky poľnohospodárov

Kapitola 5

Záver

Teoretická časť tejto diplomovej práce sa snaží systematicky zhrnúť poznatky o klasických regresných modeloch ako aj o ich zovšeobecneniach tak, aby čitateľovi podala ucelenú kompaktnú informáciu. Dôraz je smerovaný na problematiku vplyvu chybovej zložky regresného modelu a situácie, keď sú porušené predpoklady kladené na chyby. Teoretická časť poskytuje dostatok informácií pre orientáciu v danej problematike a samostatné modelovanie a vytváranie predikcií.

Vyústením diplomovej práce je aplikácia nadobudnutých znalostí v praktickej časti, v ktorej je názorne a detailne popísané budovanie Boxových–Jenkinsových modelov. Vývoj modelov prebiehal na časových radoch z praxe. Je nutné podotknúť, že sa jedná o reálne nestacionárne finančné časové rady. Táto voľba dát ukazuje použiteľnosť zvolených metód v praxi. V práci je názorne ukázaná kompletná výstavba modelov ARMA, ARIMA a SARIMA, ktoré prezentujú základnú bázu tejto metodológie.

Ďalej bol v práci porovnaný prístup eliminácie systematického trendu a následného budovania stochastického modelu s prístupom stochastického modelovania celého časového radu, vrátane trendu. Porovnanie týchto dvoch prístupov bolo vykonané na základe ďalších získaných hodnôt časového radu, ktoré medzičasom pribudli. V tomto porovnaní obstál výrazne lepšie čiste stochastický prístup. Lepší výsledok stochastického modelu bol zrejme spôsobený zmenou trendu v časovom rade. K lokálnym zmenám (ktoré nie sú v súlade s dlhodobým vývojom) trendu dochádza vo finančných časových radoch pomerne často. Preto pokiaľ pri modelovaní finančných časových radov nemáme k dispozícii expertnú znalosť o pravdepodobnom nasledujúcom vývoji trendu, je lepšie nezaoberať sa elimináciou trendu ale modelovať rad iba stochasticky.

Literatúra

- [1] Anděl J.: *Statistické metody*, MATFYZPRES, Praha, 2007.
- [2] Anděl, J.: *Statistická analýza časových řad*, SNTL, Praha, 1976.
- [3] Anděl J.: *Základy matematické statistiky*, MATFYZPRES, Praha, 2007.
- [4] Box, G.E.P.; Jenkins, G.M.: *Time Series Analysis Forecasting and Control*, Holden-Day, San Francisco, 1976.
- [5] Cipra, T.: *Analýza časových řad s aplikacemi v ekonomii*, SNTL/ALFA, Praha, 1986.
- [6] Cipra, T.: *Finanční ekonometrie*, Ekopress, Praha, 2008.
- [7] Fomby, T.B.; Hill, R.C.; Johnson, S.R.: *Advanced Econometric Methods*, Springer-Verlag, Berlin, 1984.
- [8] Goldfeld, S.M.; Quandt, R.E.: *Nonlinear Methods in Econometrics*, North Holland Publishing Company, London 1972.
- [9] Jenkins, G.M.: *Practical experiences with modelling and forecasting time series*, GJP, Jersey, 1979.
- [10] Mincer, J.A.: *Schooling and Earnings*, Columbia University Press, New York, 1974.
- [11] Shumway, H.R.; Stoffer, D.S.: *Time Series Analysis and Its Applications: With R Examples*, Springer, New York 2010.
- [12] Zvára, K.: *Regresní analýza*, Academia, Praha, 1989.

Zoznam tabuliek

3.1	Distribučné funkcie rozdelenia peakov za platnosti homoskedasticity .	50
4.1	Spočítané hodnoty Akaikeho informačného kritéria pre ARMA(p,q) modely časového radu depozitných zostatkov eliminovaného o systematický trend	76
4.2	Hodnoty Akaikeho informačného kritéria pre modely ARMA(p,q) diferencovaného časového radu depozitných zostatkov	85

Zoznam obrázkov

4.1	Skúmané časové rady.	71
4.2	Časový rad denných depozitných zostatkov	72
4.3	Trendy premietnuté do časového radu depozitných zostatkov	75
4.4	Nový časový rad obdržaný elimináciou trendu	75
4.5	Autokorelačná (ACF) a parciálna autokorelačná (PACF) funkcia nového časového radu s eliminovaným polynomiálnym trendom	76
4.6	Histogram a Q–Q graf reziduí modelu AR(4)	77
4.7	Diagnostika modelu AR(4)	78
4.8	Reziduá ARMA(2,1) a ARMA(4,1)	79
4.9	Diagnostika ARMA(2,1)	80
4.10	Diagnostika ARMA(4,1)	81
4.11	Časový rad s eliminovaným trendom s predikciami podľa ARMA(2,1) a 95% konfidenčnými intervalmi	82
4.12	Stacionarizovaný časový depozitných zostatkov s 10 predikciami budúcich hodnôt podľa AR(4)	82
4.13	Predpoveď depozitných zostatkov pre nasledujúcich 10 pracovných dní obdržaná kombináciou trendu a ARMA procesu	83
4.14	Závislosť úrovne a variability v priebehu časového radu	84
4.15	Diferencie 1. rádu	84
4.16	Autokorelačná a parciálna autokorelačná funkcia diferencií depozitných zostatkov	85
4.17	Histogramy reziduí modelov ARMA(p,q) aplikovaných na rad diferencií denných depozitných zostatkov	86
4.18	Q–Q grafy reziduí ARMA(p,q) modelov aplikovaných na rad diferencií denných depozitných zostatkov	87
4.19	Autokorelačné a parciálne autokorelačné funkcie reziduí a štvorcov reziduí modelu ARMA(5,3)	89
4.20	Diagnostika modelu ARIMA(5,1,3) aplikovaného na denné depozitné zostatky	90
4.21	Predikcie modelu ARIMA(5,1,3)	91
4.22	Časový rad denných depozitných zostatkov spolu s predikciami zostrojenými 1.čiste pomocou Boxovej–Jenkinsovej metodológie (červenou farbou), 2. elimináciou trendu a následným použitím Boxovej Jenkinsovej metodológie (modrou farbou) a 3. skutočné budúce hodnoty radu (zelenou farbou)	91

4.23	Časový rad pohľadávok na prevádzkovom financovaní	92
4.24	Stacionarizovaný časový rad pohľadávok	93
4.25	Autokorelačná a parciálna autokorelačná funkcia diferencovaného radu	93
4.26	Histogram a Q–Q graf diferencovaného časového radu	94
4.27	Diagnostika ARIMA(0,1,0)	95
4.28	Časový rad pohľadávok s predikciami	96
4.29	Mesačný časový rad depozitných zostatkov poľnohospodárov	96
4.30	Autokorelačná a parciálna autokorelačná funkcia depozitných zostatkov poľnohospodárov	97
4.31	Diferencovaný rad depozitných zostatkov poľnohospodárov	97
4.32	Diferencovaný rad depozitných zostatkov poľnohospodárov po sezónnej diferenciácii	98
4.33	Korelačná a parciálna korelačná funkcia stacionarizovaného radu . . .	98
4.34	Pôvodný časový rad s predikciou na nasledujúce dva roky	100
4.35	Diagnostika modelu SARIMA(4, 1, 3) \times (0, 1, 0) ₁₂ aplikovaného na depozitné zostatky poľnohospodárov	101